Nonparametric Methods in Financial Time Series Analysis

Seok Young Hong

Faculty of Mathematics and St Catharine's College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2018

Nonparametric Methods in Financial Time Series Analysis

Seok Young Hong
Faculty of Mathematics and St Catharine's College
University of Cambridge

The fundamental objective of the analysis of financial time series is to unveil the random mechanism, i.e. the probability law, underlying financial data. The effort to identify the truth that governs the observations involves proposing and estimating reasonable statistical models that well explain the empirical features of data. This thesis develops some new nonparametric tools that can be exploited in this context; the efficacy and validity of their use are supported by computational advancements and surging availability of large/complex ('big') data sets.

Chapter 1 investigates the conditional first moment properties of financial returns. We propose multivariate extensions of the popular Variance Ratio (VR) statistic, aiming to test linear predictability of returns and weak-form market efficiency. We construct asymptotic distribution theories for the statistics and scalar functions thereof under the null hypothesis of no predictability. The imposed assumptions are weaker than those widely adopted in the literature, and in our view more credible with regard to the underlying data generating process we expect for stock returns. It is also shown that the limit theories can be extended to the long horizon and large dimension cases, and also to allow for a time varying risk premium. Our methods are applied to CRSP weekly returns from 1962 to 2013; the joint tests of the multivariate hypothesis reject the null at the 1% level for all horizons considered.

Chapter 2 is about nonparametric estimation of conditional moments. We propose a local constant type estimator that operates with an infinite number of conditioning variables; this enables a direct estimation of many objects of econometric interest that have dependence upon the infinite past. We show pointwise and uniform consistency of the estimator and establish its asymptotic nomality in various static and dynamic regressions context. The optimal rate of estimation turns out to be of logarithmic order, and the precise rate depends on the Lambert W function, the smoothness of the regression operator and the dependence of the data in a non-trivial way. The theories are applied to investigate the intertemporal risk-return relation for the aggregate stock market. We report an overall positive risk-return relation on the S&P 500 daily data from 1950-2017, and find evidence of strong time variation and counter-cyclical behaviour in risk aversion.

Lastly, Chapter 3 concerns nonparametric volatility estimation with high frequency time series. While data observed at finer time scale than daily provide rich information, their distinctive empirical properties bring new challenges in their analysis. We propose a Fourier domain based estimator for multivariate ex-post volatility that is robust to two major hurdles in high frequency finance: asynchronicity in observations and the presence of microstructure noise. Asymptotic properties are derived under some mild conditions. Simulation studies show our method outperforms time domain estimators when two assets with different liquidity are traded asynchronously.

Preface

Whether one would be able to *understand the underlying dynamics of financial data* has long been an extremely alluring, yet a notoriously difficult and controversial, issue since Bachelier (1900) first formally brought the topic. It covers many fundamental questions in financial econometrics. Examples include whether the asset returns are predictable or how one could explain/forecast the evolution of their volatility. With a hope to add some contribution towards providing explanations to this long-standing inquiry, this dissertation develops several new nonparametric methods for time series modelling of financial data.

Financial data series of different assets in different markets are influenced − potentially heavily − by different events, news, time periods and information. Nonetheless, empirical research reveals that to our surprise the "seemingly random variations of asset prices do share some non-trivial statistical properties," Cont (2001). Those common features, widely referred to as the stylized facts, allow some unified statistical analyses of financial time series possible and make them worthwhile. In time series analysis, one first postulates a reasonable statistical model, taking into account of the stylized facts and other empirical properties of the dataset under consideration. To check if the model well-embodies the underlying true dynamics, investigation on the validity of the proposed model is often involved. Next, statistical estimation is then made with observed samples to single out the most plausible one within the proposed model.

Nonparametric approaches, now widely perceived as providing useful and flexible tools for modelling financial returns, have been extensively exploited in both contexts. In Chapter 1 we present several multivariate variance ratio statistics. They can be used to test the random walk model and hence the weak form efficient market hypothesis. Chapter 2 considers a class of time series regression models where the regressor takes values in a sequence space. The proposed nonparametric estimation method can be used in estimating the regression function conditional upon the infinite past, thereby being applicable to many important problems in finance and economics. Lastly, in Chapter 3 we propose a volatility estimator for high frequency financial time series. The estimator, based upon the Fourier-Malliavin theory, shows its competence in dealing with both issues of non-synchronously observed returns and the presence of microstructure noise.

This dissertation owes much to many people. First and foremost, I would like to express my sincere gratitude to my thesis advisor Oliver Linton for his advice, encouragement, academic and financial support, and patience throughout my PhD. I have learned enormously from him, and his guidance has been vital in writing this dissertation. I would also like to thank Alexei Onatski very much; he has always been kind and thoughtful, and provided warm encouragement. I am indebted to Philippe Vieu for his hospitality when I visited Tolouse in November 2013, and for kindly sparing his time to respond to many tedious questions I had time to time since then. I also thank Alessio Sancetta for his kind help and advice. Thanks also goes to Richard Nickl for his support and understanding when I started my PhD and thereafter.

Looking back, I would like to thank Andrew Walden, Pat Altham, Thomas Østergaard Sørensen, and Alastair Young, from whom I gained a strong motivation to pursue this research degree, a journey that has been rewarding and enlightening. During my PhD, I was extremely fortunate to learn from and collaborate with Mikhail Lifshits, Oliver Linton, Alexander Nazarov, Sujin Park and Hui Jun Zhang. I greatly appreciate their advice and helpful discussions.

I would like to thank both Statistical Laboratory and the Faculty of Economics. I learned a lot from attending the weekly Econometrics Workshop/Seminar. In particular, I was given a chance to present preliminary versions of each chapter in this dissertation in March 2014, 2015, 2017 and in June 2015, and received many helpful comments - especially from Alexei Onatski and Jeroen Dalderop. I also thank St Catharine's College for its pastoral and financial support; a substantial portion of this dissertation was written in its Shakeshaft Library.

I am indebted to Si for standing by me throughout and for always making me smile. The entire journey has been much more enjoyable and pleasant with her presence. I also thank all the friends I met in Cambridge, with whom I have had so much joyful time and shared many good memories. I am most grateful to my beloved family, in particular, to my parents for always being supportive and having trust in me. Simple words cannot fully express how thankful and appreciative I am to my parents, to whom I dedicate this thesis.

Chapter 1 is an edited version of joint work with Oliver Linton and Hui Jun Zhang published in the *Journal of Financial Econometrics* as Hong, Linton and Zhang (2017). Chapter 2 is joint work with Oliver Linton; it is a slightly modified version of what has been submitted to the *Journal of Econometrics*. The latest preprint of the paper is available in SSRN as Hong and Linton (2018). Lastly, Chapter 3 is an edited version of joint work with Oliver Linton and Sujin Park published in the *Journal of Econometrics* as Park, Hong and Linton (2016), which is an extensive update of Park (2011, Chapter 2).

Regarding notations, we define $a_n \simeq b_n$ by $a_n = b_n + o(1)$, and $c_n \sim d_n$ by equivalence of order between the two sequences c_n and d_n. Also, $f \preceq g$ means there exists some constant $c > 0$ such that $\lim_{n\to\infty} f(n)/g(n) \leq c$; similarly for $\succeq$. We take $a \wedge b$ and $a \vee b$ to mean the smaller and bigger value of the scalars a and b, respectively. The operator $\Delta(\cdot)$ acting on a time series process or a function thereof is the differencing operator; for example, $\Delta t_j = t_j - t_{j-1}$ and $\Delta g(t_j) = g(t_j) - g(t_{j-1})$. We denote by $\Longrightarrow$ the convergence in distribution, and $\Longrightarrow^{stably}$ the stable convergence in law, see Aldous and Eagleson (1978) or Jacod (1997) for rigorous definition of the latter. The term 'stationarity' is taken to mean 'strict stationarity' in Chapters 1 and 2, and 'stationarity in wide sense' in Chapter 3. The matrix norm $\|.\| = \|.\|_2$ in Chapter 3 is understood as the Frobenius norm unless specified otherwise. The square bracket notation $[\cdot,\cdot]$ for a stochastic process is taken to mean its quadratic variation. Throughout, C (or C', C'' and similar variants) refers to some generic constant that may take different values in different places unless defined specifically otherwise. All other notations are defined within the relevant texts.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

Cambridge, England
September 2018

Seok Young Hong
University of Cambridge

Contents

Chapter 1

Predictability of Asset Returns: Multivariate Variance Ratio Tests[1]

Conditional first moment properties of financial time series have been extensively studied in the financial econometrics literature. Those studies were largely driven by the curiosity to test the *random walk nature of price changes*, or more generally, the *predictability of asset returns*. It is an important question in modern financial research that offers insight into the issue of *market efficiency*.

Denote by X_t the d-dimensional vector process of asset returns and let $\widetilde{X}_t = X_t - \mu$, where $\mu = E(X_t)$ for all t. In this chapter, we aim to test the (weak form) Efficient Markets Hypothesis (henceforth for simplicity EMH) and quantifying and signing departures from this hypothesis. According to Fama (1970), this is that "risk adjusted stock returns are unforecastable using past prices". "Prices" are usually taken to mean just a sequence of past prices for the asset in question, but the spirit of this hypothesis should allow the past history of other assets not to matter either. Regarding the risk adjustment, we shall assume that the risk premium is constant, unknown, and is denoted by μ. In Hong, Linton and Zhang (2015) (henceforth often referred to as "the working paper version" of this chapter), we have extended the theory to the case where μ_t is time varying and depends on multiple unknown quantities.

One interpretation of the EMH is to assume that the risk adjusted return process satisfies

$$E(\widetilde{X}_t|\mathcal{F}_{t-1}) = 0, \tag{1.1}$$

where $\mathcal{F}_t$ denotes the past history of the prices of all the assets. This is a stronger assumption than that returns are uncorrelated with the past of all prices, i.e.,

$$E(\widetilde{X}_{it}\widetilde{X}_{jt-k}) = 0 \tag{1.2}$$

[1]Helpful comments of Xiaohong Chen, Greg Connor, Richard Davis, Tassos Magdalinos, Alexei Onatski, Andrew Patton, Hashem Pesaran, Peter Phillips, Peter Robinson, Mark Salmon, Allan Timmerman, and Wei Biao Wu are gratefully acknowledged.

for all $i, j = 1, \ldots, d$ and for all $k \neq 0$, which itself is a stronger assumption[2] than that returns are uncorrelated with their own past, i.e.,

$$E(\widetilde{X}_{it}\widetilde{X}_{it-k}) = 0 \tag{1.3}$$

for all i and for all $k \neq 0$, which is what is adopted in Lo and MacKinlay (1988) (and referred to as the Random Walk 3 (RW3) in Campbell, Lo, and MacKinlay (1997) and in much subsequent work). RW3 has the advantage that if one rejects it, then one rejects the martingale hypothesis; on the other hand, if one does not reject RW3 then one can't conclude that the martingale hypothesis is valid.[3] Throughout this chapter, we work with at least the multivariate uncorrelatedness hypothesis (1.2). We also develop a theory based on the stronger martingale difference assumption (1.1), because the additional regularity conditions can be stated very simply.

In this chapter, we propose several *multivariate* extensions to the variance ratio statistic, a popular test statistic widely adopted in the empirical finance literature. Those statistics can be used in "testing" the weak form EMH and for measuring the direction and magnitude of departures from this hypothesis. Asymptotic distributions of the statistics and scalar functions thereof are derived, under the null hypothesis that returns are unpredictable after a constant mean adjustment. The methodology is applied to weekly returns for Center for Research in Security Price (CRSP) size-sorted portfolios from 1962 to 2013 in three subperiods. We find evidence of a reduction of linear predictability in the most recent period, for small and medium cap stocks, but we still reject the multivariate null hypothesis in the most recent period. The main findings are not substantially affected by allowing for a common factor time varying risk premium.

1.1 Introduction

It is fair to say that the profession is divided on the evidence regarding the EMH. Authors like Fama (1970, 2013), Malkiel (2015), and Ross (2002) argue that rejections of the EMH are: small, not scalable, fleeting, statistically suspect, and not realizable profit opportunities. Furthermore, Fama has emphasized the joint hypothesis problem whereby one must measure abnormal returns relative to a market equilibrium return that provides compensation for bearing risk, so that statistical rejections of the hypothesis are potentially due instead to rejection of the assumed market equilibrium return. On the other hand, authors like Shiller (2013), Kahneman and Tversky (2000) and others have argued

[2] This is not quite correct, since the martingale hypothesis only requires $E|X_t| < \infty$, whereas the autocovariance of a stationary process requires $EX_t^2 < \infty$ in order to be well defined in general.

[3] We note that there are many tests of the martingale hypothesis that make use of more information, Hong and Lee (2005) and Escanciano and Velasco (2006), and thereby obtain power against a larger class of alternatives.

that market participants are irrational, behave irrationally, and that their interaction produces excess volatility in asset returns relative to fundamentals. Grossman and Stiglitz (1980) argue that even if market participants are fully rational, if information acquisition is costly, then prices cannot perfectly reflect the information which is available, since if it did, those who spent resources to obtain it would receive no compensation, leading to the conclusion that an informationally efficient market is impossible. King (2016) emphasizes the "radical uncertainty" issue, whereby the future states of the world are not perfectly knowable and formal concepts such as probability distributions and expectations that are core to EMH are of limited use. These are just some of the many varied and nuanced points of view on this subject.

Our purpose is not to provide definitive evidence on this hypothesis one way or another; instead we focus on some methodological issues. As Robert Shiller says in his Nobel lecture: "Ultimately, the question in reconciling the apparently conflicting views comes down to that of constructing the right statistical tests." We contribute to this by investigating a celebrated class of tests of this hypothesis, which we think have been wrongly applied, and making some modest proposals to improve best practice with regard to their use.

Variance ratio statistics (Lo and MacKinlay (1988) and Poterba and Summers (1988)) are widely used in empirical finance as a way of testing the EMH and to measure the degree and (cumulative) direction of departures from this hypothesis in financial time series. Indeed, this work has been extremely influential in understanding predictability in asset prices and in measuring market quality. A key advantage of this methodology relative to say Box-Pierce statistics is that variance ratios give information about the direction of departures from the null hypothesis that can be interpreted in meaningful economic terms (i.e., momentum versus contrarian), so that the analysis does not reduce to yes/no decision-making on an uninformative test statistic. A lot of empirical work followed immediately after the seminal contributions. Lo and MacKinlay (1988) presented evidence regarding predictability of the US stock market. They concluded that the EMH was soundly rejected in weekly US stock market returns based on their standard errors. The graduate textbook Campbell, Lo, and MacKinlay (1997), henceforth CLM, presents variance ratios for weekly value weighted and equal weighted CRSP indexes and five size sorted portfolios over the period 1962-1994; they argue that the EMH is strongly rejected based on their standard errors, although they find that the magnitude of the violation is less in the later subperiod 1978-1994. On the other hand, Cochrane (2001, p388) writing only four years later argues that: "daily, weekly, and monthly stock returns are close to unpredictable".[4] One important recent direction for this methodology is in "high frequency" settings, i.e., intraday, where it has informed the debate on the

[4]He then emphasized the more recent work that had shown that low frequency returns (business cycle and longer) are predictable from dividend price ratio and term premium variables.

evolution of "market quality" in the US stock market. Castura, Litzenberger, Gorelick, and Dwivedi (2010) investigate trends in market efficiency in Russell 1000/2000 stocks over the period 1 January 2006 to 31 December 2009. Based on evidence from intraday variance ratios (they look at 10:1 second variance ratios as well as 60:10 and 600:60 second ratios), they argue that markets have become more efficient at the high frequency over time. Chordia, Roll, and Subrahmanian (2011) compared intraday variance ratios over the period 1993-2000 with the period 2000-2008 and found that the hourly to daily variance ratios of NYSE listed stocks came closer to the EMH predicted values on average in the second period.[5] One interpretation of these studies is that the computerized trading systems that now dominate equity markets have improved the functioning of those markets. Hasbrouck (2015) has recently used variance ratios to measure high frequency volatility in quoted prices, which also relates to this question. Finally, given that variance ratios are a standard measure of market quality, they are often used in cross sectional or panel data regressions as dependent variables, see for example O'Hara and Ye (2009). In short, variance ratios are the *de facto* measure of predictability/market efficiency that is adopted universally by financial empiricists. It is important therefore that this class of tests be given a firm foundation.

There have been some criticisms of the univariate variance ratio methodology as a test of uncorrelatedness. Specifically, it is not consistent against all (fixed of given order) alternatives unlike the Box-Pierce statistics. It is a linear functional of the autocorrelation function and so provides no new information relative to that. It seems like a redundant test. Faust (1992) provides some intellectual credibility: he shows that they can be given a likelihood ratio test interpretation and are optimal against certain alternatives of the mean reverting type. In that sense they are similar to the Durbin-Watson test. The advantage of the variance ratio over the Box-Pierce statistic is that it gives some sense of the direction of predictability, which is lost in the Box-Pierce or other portmanteau tests. Hillman and Salmon (2007) have argued that the variance ratio (actually the related variogram) is better suited to irregularly spaced data and some kinds of nonstationarity than correlogram tests. Finally, there is a lot of work on improving the finite sample performance (size and power) of both Box-Pierce statistics and variance ratio statistics, see for example Kim, Nelson, and Startz (1991) and Kan and Wang (2010). See Charles and Darné (2009) for a recent review of this methodology and its application.

We make several contributions. First, we develop a multivariate methodology. Many tests of the EMH have been carried out using the univariate variance ratio approach, that is, conducted one asset at a time. This chapter proposes a methodology for multivariate variance ratio tests. The rationale for the test is roughly the following. Suppose that the EMH hypothesis is not rejected for asset i based on univariate variance ratio tests. Suppose however that returns on i are predicted by lags of some other variable.

[5] See also Sheppard (2013) for some theoretical results using a continuous time framework.

A univariate test could fail to detect this violation of the EMH, although a multivariate test could detect it. This generic argument about the efficacy of multivariate versus univariate methods is widely accepted. There is a lot of work on multivariate portmanteau statistics, i.e., generalizations of the Box-Pierce statistic to multivariate time series, see for example Chitturi (1974) and Hosking (1981). The variance ratio statistics convey directional information about cross-autocorrelations beyond that contained in the portmanteau statistics, that is, in the case of a violation of the hypothesis they give some sense of the direction of departure. The univariate variance ratios describe the behaviour of the asset variances, whereas the multivariate statistics also measure the behaviour of the cross correlations and their cumulative direction. This could be important for momentum based portfolio trading strategies, for example.

Second, we propose an alternative distribution theory and standard errors (*heteroskedasticity and leverage consistent HLCM*) than are usually adopted (i.e., in the univariate case). The limiting distribution established in Lo and MacKinlay (1988, Theorem 3) and repeated in CLM (and so used in most empirical studies) for the univariate variance ratio statistics is incorrect under their stated assumptions H1-H4 (i.e., RW3).[6] The correct distribution would be much more complicated and would depend on a long run variance that may be hard to estimate well. Either one makes additional assumptions to ensure that the variance is as claimed, which is what we propose below, or one has to use more complicated inference methods based on long run variance estimation, Newey and West (1987), or self normalization, Lobato (2001). In fact, the omitted condition appears quite innocuous, so their essential approach seems correct. However, we think that the no-leverage assumption (Lo and MacKinlay's H4) is untenable, empirically. Although this latter condition is satisfied by GARCH volatility processes with symmetrically distributed innovations, it is not satisfied by volatility processes that allow for leverage effects such as the GJR GARCH process or the Nelson's EGARCH process, and it is not even satisfied by standard GARCH volatility processes where the innovation is asymmetric. The statistical value of the restriction is that it simplifies the standard error calculation, although, as we show, the standard errors that allow for violations of this condition do not entail an inordinate increase in computation or complexity. Essentially, Lo and MacKinlay (1988) imposed an unnecessary assumption but fail to impose a necessary one. We propose modified assumptions that still preserve the possibility of simple inference methods but allow for leverage effects. Specifically, we establish the asymptotic distribution of our statistics under two sets of assumptions: (a) a stationary martingale difference hypothesis with fourth unconditional moments; (b) uncorrelatedness as in Lo and MacKinlay (1988) and with an additional uncorrelatedness condition on the products of returns but without the additional no-leverage condition. The asymptotic variance is the same under our two different sets of assumptions but is different from that contained in Theorem 3 of

[6]It makes use of the CLT developed by White and Domowitz (1984) and used by many others.

Lo and MacKinlay (1988) (and used in much subsequent empirical work). We remark that their theory essentially imposes that the sample autocorrelations are asymptotically uncorrelated, which can lead to inappropriate standard errors and p-values. This fact has been long appreciated in the time series literature, see for example Dufour and Roy (1985); Francq, Roy, and Zakoian (2005) have provided a comprehensive theory for Box-Pierce statistics under stationarity and mixing conditions.

We propose a simple analogue method for conducting inference that does not require the selection of a bandwidth parameter. We note that the evidence about predictability of asset returns in a large number of papers has been based on the Lo and MacKinlay (1988) standard errors, which we argue should be replaced by standard errors that rely on weaker and more plausible assumptions.[7] We show that in practice the standard errors can make a difference, especially when the time series is short (such as when stationarity is of concern).

Third, we extend our null hypothesis limit theory to the long horizon and large dimension cases. We derive the null limiting distribution of the studentized statistics under the increasing horizon framework, and show that asymptotic normality holds albeit with a slower rate of convergence, extending the univariate results of Chen and Deo (2006). We also establish the same result for the average scalar variance ratio statistic in the case where the horizon is fixed but the dimensions of the vector time series increase with sample size. Fourth, we also establish the asymptotic properties of our statistic under several plausible alternative models including a multivariate Muth (1960) fads model and the recently developed bubble process of Phillips and Yu (2011). These alternatives yield quite different predictions regarding the long run value of the variance ratio statistics.

Finally, we apply our methods to weekly returns for CRSP size-sorted portfolios from 1962 to 2013 in three subperiods 1962-1978, 1978-1994 and 1994-2013; the first two subperiods correspond to the data used in CLM. We show that the degree of inefficiency has reduced over the most recent period, and in some cases this improvement is statistically significant. Specifically, the univariate tests do not reject the null hypothesis for medium or large stocks in the most recent period. However, the multivariate tests do reject, albeit with a lower significance level. We have also extended our analysis to allow for a time varying risk premium, but find that the main empirical results are sustained, and we omit these results here.[8] This evidence is presented based on our HLCM standard errors that are robust to leverage effects as well as heteroskedasticity. We also show that the degree of asymmetry in the dependence structure has reduced, although it is still statistically

[7] At the current count there were 4132 google citations of that paper.

[8] In the working paper version of this chapter, we extend the theory to allow for a time varying risk premium in two ways. One approach is to fit an observable common factor regression and compute our statistics from the residuals. The second approach is to fit explicitly a nonparametric trend model, which we also allow to vary across different "regimes" (such as days of the week), to each series, and then to compute our statistics from the residuals. We show that with minor additional conditions our distribution theory and inference method carry over to this case.

significant. We further investigate the variance ratios at the long horizon. Simulation experiments indicate that our variance ratio tests are reliable, and powerful against some alternatives.

There is a substantial literature on testing for nonlinear predictability using information beyond the simple autocorrelations, see for example Hong (2000), Hong and Lee (2005), Escanciano and Velasco (2006), and Phillips and Jin (2014). There is also a literature that emphasizes structural breaks and rolling window analysis, see for example Lo (2005) and Pesaran and Timmermann (2007). Finally, there is a large literature on "predictive regressions" using long horizons and covariates such as dividend price ratios, see Phillips (2015). Our methodology and application hopefully complements this vast body of research.

In Section 1.2 we introduce the multivariate ratio population statistics in various forms. In Section 1.3 we introduce the estimators, while in Section 1.4 we present the main central limit theorem and inference methods. In Section 1.5 we consider a number of alternative hypotheses, while in Section 1.6 we discuss the large dimensional case. We perform a small simulation study in Section 1.7. In Section 1.8 we present our application, while Section 1.9 concludes. The appendix contains the proofs of all results.

1.2 Multivariate Variance Ratios

Recall the discrete time series $X_t \in \mathbb{R}^d$ defined in the beginning of this chapter. For expositional purposes we shall suppose in this section that X_t is stationary ergodic; formal assumptions regarding the data are given below in Section 1.3.

We next define the population versions of the multivariate variance ratios. Let $X_t(K) = X_t + X_{t-1} + \ldots + X_{t-K+1}$ for each K, and define the following population quantities:

$$\Sigma = \mathrm{var}(X_t) = E(\widetilde{X}_t \widetilde{X}_t^\intercal) \tag{1.4}$$

$$D = \mathrm{diag}\left\{E(\widetilde{X}_{1t}^2), \ldots, E(\widetilde{X}_{dt}^2)\right\} \tag{1.5}$$

$$\Sigma(K) = \mathrm{var}(X_t(K)) = E((X_t(K) - KE(X_t))\,(X_t(K) - KE(X_t))^\intercal) \tag{1.6}$$

$$\Gamma(j) = \mathrm{cov}(X_t, X_{t-j}) = E(\widetilde{X}_t \widetilde{X}_{t-j}^\intercal) \tag{1.7}$$

$$R(j) = \Sigma^{-1/2}\Gamma(j)\Sigma^{-1/2} \tag{1.8}$$

$$R_L(j) = \Gamma(j)\Sigma^{-1} \quad ; \quad R_R(j) = \Sigma^{-1}\Gamma(j) \tag{1.9}$$

$$Rd(j) = D^{-1/2}\Gamma(j)D^{-1/2} \tag{1.10}$$

for $j = 0, \pm 1, \ldots$. Here, $A^{1/2}$ denotes a symmetric square root of a symmetric matrix A. We shall assume that Σ is strictly positive definite.

1.2.1 Two Sided Variance Ratios

Under condition (1.2), the variance covariance matrices obey the scaling law $\text{var}(X_t(K)) = K\text{var}(X_t)$, where K is some positive integer, from which we may obtain a number of different variance ratio statistics. These will have different merits and drawbacks depending on the purpose to which the estimation/testing is directed.

We define the two sided matrix normalized multivariate ratio (population) statistic as

$$\mathcal{VR}(K) = \text{var}(X_t)^{-1/2}\text{var}(X_t(K))\text{var}(X_t)^{-1/2}/K. \tag{1.11}$$

Clearly, under the null hypothesis (1.2) we should have $\mathcal{VR}(K) = I_d$. Under the generic (stationary) alternative hypothesis we have

$$\mathcal{VR}(K) = I + \sum_{j=1}^{K-1}\left(1 - \frac{j}{K}\right)\left(R(j) + R(j)^{\intercal}\right), \tag{1.12}$$

which is a symmetric matrix. The off-diagonal elements should be zero under the null hypothesis of no predictability. Both representations (1.11) and (1.12) can be used as the basis for estimation.[9]

An alternative multivariate normalization is given by

$$\mathcal{VR}a(K) = \text{var}(X_t(K))\text{var}(X_t)^{-1}/K,$$

which can likewise generically be written

$$\mathcal{VR}a(K) = I + \sum_{j=1}^{K-1}\left(1 - \frac{j}{K}\right)\left(R_L(j) + R_R(j)^{\intercal}\right). \tag{1.13}$$

This has a regression interpretation, see Chitturi (1974) and Wang (2003, page 62). Note that $\mathcal{VR}(K) = I$ if and only if $\mathcal{VR}a(K) = I$. We shall not say anything further about this quantity $\mathcal{VR}a(K)$. Some discussion is given in the working paper version, Hong, Linton and Zhang (2015).

A third quantity is the diagonally normalized variance ratio

$$\begin{aligned}\mathcal{VR}d(K) &= D^{-1/2}\text{var}(X_t(K))D^{-1/2}/K &(1.14)\\ &= Rd(0) + \sum_{j=1}^{K-1}\left(1 - \frac{j}{K}\right)\left(Rd(j) + Rd(j)^{\intercal}\right), &(1.15)\end{aligned}$$

where $Rd(0) = D^{-1/2}\Gamma(0)D^{-1/2}$ is the $d \times d$ contemporaneous correlation matrix. Under

[9] One can interpret the variance ratio matrix as a (scalar) affine transformation of the least squares closest value of $\mathcal{R}$ in an approximating model for the autocorrelations of the form: $\mathcal{R}(j) = \left(1 - \frac{j}{K}\right)\mathcal{R}$, $j = 1, \ldots, K$ and $\mathcal{R}(j) = 0$ for $j > K$.

the null hypothesis that the series is uncorrelated, we should have $\mathcal{VR}d(K) = Rd(0)$ the contemporaneous correlation matrix, whose off-diagonal elements are unrestricted by the null hypothesis. The diagonal elements of $\mathcal{VR}d(K)$ correspond to the univariate variance ratio statistics, while the off-diagonal elements provide information about the cumulative cross-dynamics between the assets. Note that if $\mathcal{VR}(K) = I$, then $\mathcal{VR}d(K)_{ii} = 1$ for all i, but not vice versa. This suggests that if one rejects a univariate test then one would reject the multivariate test but not necessarily vice versa. Specifically, suppose that X_t are iid but $X_{1t} = X_{2,t-1}$ then the univariate tests would fail but the multivariate one would not.

1.2.2 One Sided Variance Ratios

In the univariate case, the variance ratio process and the autocorrelation function contain the same information and one can recover the autocorrelation function from the variance ratio function. This is not so in the multivariate case because $\mathcal{VR}(K)$ and $\mathcal{VR}d(K)$ are both symmetric matrices whereas the autocorrelation function $Rd(j)$ is not necessarily symmetric. In fact, one can only recover $Rd(\cdot) + Rd(\cdot)^\intercal$ or $R(\cdot) + R(\cdot)^\intercal$ from the variance ratio functions $\mathcal{VR}d(\cdot)$ and $\mathcal{VR}(\cdot)$. This means that information about lead lag relations are eliminated. Instead we propose the following quantities:

$$\mathcal{VR}_+(K) = I + 2\sum_{j=1}^{K-1}\left(1 - \frac{j}{K}\right) R(j)$$

$$\mathcal{VR}d_+(K) = Rd(0) + 2\sum_{j=1}^{K-1}\left(1 - \frac{j}{K}\right) Rd(j),$$

and the negative counterparts $\mathcal{VR}_-(K) = \mathcal{VR}_+^\intercal(K)$ and $\mathcal{VR}d_-(K) = \mathcal{VR}d_+^\intercal(K)$, which have the property that: $\mathcal{VR}(K) = (\mathcal{VR}_+(K) + \mathcal{VR}_+^\intercal(K))/2$ and $\mathcal{VR}d(K) = (\mathcal{VR}d_+(K) + \mathcal{VR}d_+^\intercal(K))/2$. One can compare the two statistics, $\mathcal{VR}d_+(K)$, $\mathcal{VR}d_-(K)$, to quantify the asymmetry in lead lag effects.

1.2.3 Univariate Parameters of Interest

We discuss here some univariate parameters of interest both for statistical purposes and economic interpretability.

(1) Trace and Determinant

The determinant and trace are commonly used univariate functions of covariance matrices that feature in a lot of likelihood ratio testing literature, see for example Szroeter (1978). The trace statistic is widely used to capture the average effect of many individual variance ratios, see for example Table 2.3 in Lo and MacKinlay (1999), and Castura et al. (2010). The Generalized Variance Ratio (Anderson (2003)) statistic would be

$$\det\left(\mathcal{VR}(K)\right) = \frac{\det\left(\Sigma(K)/K\right)}{\det\left(\Sigma\right)} = \frac{\det\left(\Sigma(K)\right)}{K^d \det\left(\Sigma\right)}.$$

Cho and White (2014) Lemma 1 says that $\mathcal{VR}(K) = I$ if and only if $\det\left(\mathcal{VR}(K)\right) = 1$ and $\mathrm{tr}\left(\mathcal{VR}(K)\right) = d$, so from a statistical point of view these quantities capture the meaning of the null hypothesis.

(2) Eigenvalues

Let $\lambda_1(K) = \lambda_{\max}(K) \geq \ldots \geq \lambda_d(K) = \lambda_{\min}(K)$ denote the eigenvalues of $\mathcal{VR}(K))$ arranged in decreasing order. Under the null hypothesis, $\lambda_j(K) = 1$, $j = 1, \ldots, d$, but under the alternative hypothesis they can take any non-negative values. These quantities give univariate measures of the predictability obtainable within the series as we next show. Consider a portfolio of assets with fixed weights $w \in \mathbb{R}^d$. Denoting $\mathcal{VR}_K(z_t)$ by the univariate variance ratio of the scalar series z_t, and letting $\widetilde{w} = \Sigma^{1/2} w$ and $Y_t = \Sigma^{-1/2} X_t$, we have

$$\begin{aligned}
\mathcal{VR}_K(w^\intercal X_t) &= \mathcal{VR}_K(w^\intercal \Sigma^{1/2}\Sigma^{-1/2} X_t) = \mathcal{VR}_K(\widetilde{w}^\intercal Y_t) = \frac{\widetilde{w}^\intercal \mathcal{VR}(K; Y_t)\widetilde{w}}{\widetilde{w}^\intercal \widetilde{w}} \\
&= \frac{\widetilde{w}^\intercal \mathcal{VR}(K; X_t)\widetilde{w}}{\widetilde{w}^\intercal \widetilde{w}} \leq \lambda_1(\mathcal{VR}(K; X_t)).
\end{aligned}$$

This follows because $\mathcal{VR}(K; X_t) = \mathcal{VR}(K; \Sigma^{-1/2} X_t) = \mathcal{VR}(K; Y_t)$. This says that the largest eigenvalue of the variance ratio matrix is an upper bound on the univariate variance ratio of any portfolio with fixed ex-post weights. Likewise, the smallest eigenvalue of the variance ratio matrix provides a lower bound on the variance ratio of any portfolio with fixed weights. The weights that achieve it are given by the corresponding rescaled eigenvectors of the variance ratio matrix. Compare with Lo and MacKinlay (1999, page 258). The portfolio that gives minimal predictability corresponds to the eigenvalue $\lambda_j(K)$ that is closest to one.

(3) Global Minimum Variance

The variance ratio matrix can also tell us about other portfolios constructed from the underlying assets. The variance of the portfolio $w^\intercal X_t(K)$ is $w^\intercal \Sigma(K) w$. Denote by i the d-dimensional column vector of ones. The global minimum variance portfolio weights are $w_{mv}(K) = \Sigma(K)^{-1} i / i^\intercal \Sigma(K)^{-1} i$, which results in global minimum variance $1/i^\intercal \Sigma(K)^{-1} i$. By plotting this as a function of K one sees the variation of the least risk portfolio by horizon. This comparison does not depend on the matrix Σ so if we consider the normalized returns $Y_t(K) = K^{-1/2}\Sigma^{-1/2}(X_t(K) - K\mu)$ then the variance of $w^\intercal Y_t(K)$ is $w^\intercal \Sigma^{-1/2}\Sigma(K)\Sigma^{-1/2}w/K = w^\intercal \mathcal{VR}(K)w$ and the best portfolio is $w_{mv}(K) = \mathcal{VR}(K)^{-1}i/i^\intercal \mathcal{VR}(K)^{-1}i$ with resulting variance

$$GMV(K) = \frac{1}{i^\intercal \mathcal{VR}(K)^{-1}i}. \tag{1.16}$$

Under the null hypothesis this should be equal to $1/d$ for all K.

(4) Off-Diagonal Elements

We are also interested in several other univariate parameters based on $\mathcal{VR}d_+(K)$. First, the diagonal elements of $\mathcal{VR}d_+(K)$ correspond to the univariate variance ratio statistics. Second, the off-diagonal elements of $\mathcal{VR}d_+(K)$ provide the information about the directional lead lag pattern between the assets. Third, the differences between two corresponding off-diagonal elements of $\mathcal{VR}d_+(K)$ indicate the asymmetry in the lead lag relationships between the assets. If one of the assets is a common factor portfolio, the corresponding off-diagonal elements of $\mathcal{VR}d_+(K)$ and $\mathcal{VR}d_-(K)$ give an idea of the dynamic comovement of the asset with the common factor portfolio, which could be used in cross-sectional regression analysis.

Another parameter of interest is the average of the off diagonal elements of $\mathcal{VR}d(K)$, which is

$$CS(K) = \frac{2}{d(d-1)}\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\mathcal{VR}d_{ij}(K) = \frac{1}{d(d-1)}\{i^\intercal \mathcal{VR}d(K)i - \mathrm{tr}(\mathcal{VR}d(K))\}, \tag{1.17}$$

see Solnik (1991) and Bailey, Kapetanios, and Pesaran (2012) who consider the case of $K = 1$ and large d. Under the null hypothesis $CS(K) = CS(1)$ for all K. This measures in some average sense the cross dependence at different lags.

(5) Dynamic Momentum/Contrarian Portfolio Profit

We consider a generalization of the Lo and MacKinlay (1990) type arbitrage portfolio contrarian strategies. Specifically, consider the following portfolio weights applied to the normalized investments $Z_t = D^{-1/2}(X_t - \mu)$

$$\widetilde{w}_{it}(K) = \pm\frac{2}{d(K-1)}\sum_{j=1}^{K-1}\left(1 - \frac{j}{K}\right)\left(Z_{i,t-j} - \overline{Z}_{t-j}\right) \tag{1.18}$$

where $\overline{Z}_s = \sum_{i=1}^d Z_{is}/d = i^\intercal Z_s/d$ so that $\sum_{i=1}^d \widetilde{w}_{it}(K) = 0$. This strategy considers all the "signals": $Z_{i,t-1} - \overline{Z}_{t-1}, \ldots, Z_{i,t+1-K} - \overline{Z}_{t+1-K}$, and combines them with weights according to their lag. If the $\pm$ factor is positive, this can be considered a momentum strategy, while if it is negative, this can be considered a contrarian strategy. The total investment of the strategy at time t is $I_t(K) = \sum_{i=1}^{d} |\widetilde{w}_{it}(K)|/2$. The expected profit of this strategy is

$$\begin{aligned}
\pi_{\pm}(K) &= E\widetilde{w}_t^{\intercal}(K)Z_t = \pm\frac{2}{d(K-1)}\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)E\left[\left(Z_{t-j}-\overline{Z}_{t-j}i\right)^{\intercal}Z_t\right] \\
&= \pm\frac{2}{d^2(K-1)}\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)E\left[i^{\intercal}Z_{t-j}Z_t^{\intercal}i\right]-\left[\pm\frac{2}{d(K-1)}\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)E\left[Z_{t-j}^{\intercal}Z_t\right]\right] \\
&= \pm\frac{2}{d^2(K-1)}i^{\intercal}\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)R(j)^{\intercal}i-\left[\pm\frac{2}{d(K-1)}\mathrm{tr}\left(\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)R(j)\right)\right] \\
&= \pm\frac{1}{d^2(K-1)}i^{\intercal}\mathcal{VR}d(K)i-\left[\pm\frac{1}{d(K-1)}\mathrm{tr}\left(\mathcal{VR}d(K)\right)\right]\pm\frac{1}{K-1}\left(1-\frac{1}{d^2}i^{\intercal}R(0)i\right) \\
&= \pm\frac{2}{d^2(K-1)}\sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\left[\mathcal{VR}d_{ij}(K)-\rho_{ij}\right]\pm\frac{d-1}{d^2(K-1)}\mathrm{tr}\left(I-\mathcal{VR}d(K)\right).
\end{aligned}$$

Under the martingale hypothesis, $\pi_{\pm}(K) = 0$ for all K. This quantity weights diagonal departures and off diagonal departures similarly. If $\pi_{\pm}(K) > 0$, then the strategy should make money (in the absence of transaction costs).

1.3 Estimation of Variance Ratio Matrices

Suppose that we observe the return vectors $\{X_t, t = 1, \ldots, T\}$ equally spaced in discrete time. We may estimate the variance ratios in several ways, for example by estimating the sample covariance matrix of the K frequency data and the original observations and then forming the ratio.[10] We can alternatively explicitly use the population connection with the autocorrelation matrix process in (1.12) for example.

We estimate the population quantities by sample averages:

$$\overline{X} = \frac{1}{T}\sum_{t=1}^{T}X_t \quad ; \quad \widehat{\Gamma}(j) = \frac{1}{T}\sum_{t=j+1}^{T}\left(X_t-\overline{X}\right)\left(X_{t-j}-\overline{X}\right)^{\intercal},\ j=0,1,2,\ldots$$

$$\widehat{\Sigma}(K) = \frac{1}{T}\sum_{t=K}^{T}\left(X_t(K)-K\overline{X}\right)\left(X_t(K)-K\overline{X}\right)^{\intercal}$$

$$\widehat{\Sigma} = \widehat{\Gamma}(0) \quad ; \quad \widehat{D} = \mathrm{diag}[\widehat{\Gamma}(0)] \quad ; \quad \widehat{R}(j) = \widehat{\Sigma}^{-1/2}\widehat{\Gamma}(j)\widehat{\Sigma}^{-1/2}$$

$$\widehat{R}d(j) = \widehat{D}^{-1/2}\widehat{\Gamma}(j)\widehat{D}^{-1/2}; \quad \widehat{\mathcal{VR}}(K) = I+\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)\left(\widehat{R}(j)+\widehat{R}(j)^{\intercal}\right)$$

[10] As pointed out by Hillman and Salmon (2007) with unequally spaced data, this approach can yield a "natural" variance ratio by classifying observations on the duration since the previous trade.

$$\widehat{\mathcal{VR}}d(K) = I + \sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)\left(\widehat{R}d(j) + \widehat{R}d(j)^{\intercal}\right)$$

$$\widehat{\mathcal{VR}}^{\&}(K) = \widehat{\Sigma}^{-1/2}\widehat{\Sigma}(K)\widehat{\Sigma}^{-1/2}/K; \quad \widehat{\mathcal{VR}}_{+}(K) = I + 2\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)\widehat{R}(j).$$

Note that by construction $\widehat{\mathcal{VR}}(K)$, $\widehat{\mathcal{VR}}d(K)$, and $\widehat{\mathcal{VR}}^{\&}(K)$ are symmetric and positive semidefinite. We may also calculate the univariate quantities by analogy. For example, define the estimated ordered eigenvalues $\widehat{\lambda}_1(K) \geq \ldots \geq \widehat{\lambda}_d(K)$ of $\widehat{\mathcal{VR}}(K)$.

1.4 Asymptotic Theory and Inference

1.4.1 Regularity Conditions

We present two alternative non-nested sets of sampling assumptions, which we denote by A and MH*. Assumptions A center on the martingale difference assumption and require stationarity and ergodicity. The theory makes use of arguments presented in Hall and Heyde (1980), and applied in Phillips and Guo (2001); see Escanciano and Lobato (2009) for a review of the literature surrounding martingale based testing. Assumptions MH* are modified versions of the assumptions in Lo and MacKinlay (1988) adapted to the multivariate case and corrected for what appears to be an error; these conditions do not require stationarity although certain averages need to converge. Most treatments of variance ratios employ the Lo and MacKinlay (1988) assumption H, which includes a mixing condition and some further restriction on the structure of the higher moments (their condition H4), which purportedly implies that the sample autocorrelations are asymptotically independent.[11] In the multivariate context, their assumption H4 would be that

$$E[\widetilde{X}_{it}\widetilde{X}_{jt}\widetilde{X}_{kr}\widetilde{X}_{ls}] = 0 \text{ for all } i,j,k,l,t, \text{ and } r,s \text{ with } r < s < t. \tag{1.19}$$

This assumption rules out leverage type effects, e.g., $E[\widetilde{X}_{ir}\widetilde{X}_{is}|\widetilde{X}_{it}^2] \neq 0$, which may be important for some assets, see Nelson (1991). This assumption is not necessary for the distribution theory; imposing it (along with other conditions) would simplify the asymptotic variance to be single finite sums rather than double finite sums, but in practice this is not a big issue. We shall dispense with this assumption below, but we shall make a further assumption that appears to have been omitted by mistake from Lo and MacKinlay (1988). Namely, implicit in their analysis is that $\widetilde{X}_t\widetilde{X}_{t-j}$ is uncorrelated with $\widetilde{X}_s\widetilde{X}_{s-j}$, but this does not follow from $\widetilde{X}_t$ being an uncorrelated sequence (although it does follow if $\widetilde{X}_t$ were a martingale difference sequence).

[11] Some papers including Whang and Kim (2003) dispense with this latter assumption but maintain the mixing and moment assumption.

Define for $j, k = 0, 1, 2, \dots$:

$$\Xi_{jk} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E\left[\left(\widetilde{X}_{t-j}\widetilde{X}_{t-k}^{\intercal} \otimes \widetilde{X}_t\widetilde{X}_t^{\intercal}\right)\right] \quad ; \quad c_{j,K} = 2\left(1 - \frac{j}{K}\right)$$

$$\begin{aligned} Q(K) &= \sum_{j=1}^{K-1}\sum_{k=1}^{K-1} c_{j,K}c_{k,K}\left(\Sigma^{-1/2}\otimes\Sigma^{-1/2}\right)\Xi_{jk}\left(\Sigma^{-1/2}\otimes\Sigma^{-1/2}\right) \\ Qd(K) &= \sum_{j=1}^{K-1}\sum_{k=1}^{K-1} c_{j,K}c_{k,K}\left(D^{-1/2}\otimes D^{-1/2}\right)\Xi_{jk}\left(D^{-1/2}\otimes D^{-1/2}\right). \end{aligned}$$

We shall assume that the matrices Σ, $Q(K)$ and $Qd(K)$ are strictly positive definite. We consider the following sets of alternative assumptions:

Assumption A.

A1. *The process $\widetilde{X}_t$ is a stationary ergodic Martingale Difference sequence;*

A2. *The process $\widetilde{X}_t$ has finite fourth moments, i.e., for all i, j, k, l, $E[|\widetilde{X}_{it}\widetilde{X}_{jt}\widetilde{X}_{kt}\widetilde{X}_{lt}|] \leq C < \infty$.*

Assumption MH*.

MH1. (i) *For all t, $\widetilde{X}_t$ satisfies $E\widetilde{X}_t = 0$, $E\big[\widetilde{X}_t\widetilde{X}_{t-j}^{\intercal}\big] = 0$ for all $j \neq 0$;* (ii) *for all t, s with $s \neq t$ and all $j, k = 1, \dots, K$, $E\big[\widetilde{X}_t\widetilde{X}_{t-j}^{\intercal} \otimes \widetilde{X}_s\widetilde{X}_{s-k}^{\intercal}\big] = 0$.*

MH2. *$\widetilde{X}_t$ is α-mixing with coefficient $\alpha(m)$ of size $r/(r-1)$, where $r > 1$, such that for all t and for any $j \geq 0$, there exists some $\delta > 0$ for which $\sup_t E|\widetilde{X}_{it}\widetilde{X}_{k,t-j}|^{2(r+\delta)} < \Delta < \infty$ for all $i, k = 1, \dots, d$;*

MH3. *For all j, k, the following limits exist: $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} E[\widetilde{X}_t\widetilde{X}_t^{\intercal}] =: \Sigma < \infty$ and $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} E\big[\widetilde{X}_{t-j}\widetilde{X}_{t-k}^{\intercal} \otimes \widetilde{X}_t\widetilde{X}_t^{\intercal}\big] =: \Xi_{jk} < \infty$.*

Chen and Deo (2006) work with martingale difference sequences but also assume a no leverage condition. Francq, Roy, and Zakoian (2005) assume both stationarity and mixing in their analysis of Box-Pierce statistics. In MH* we include the additional condition (ii) $E[\widetilde{X}_t\widetilde{X}_{t-j}^{\intercal} \otimes \widetilde{X}_s\widetilde{X}_{s-k}^{\intercal}] = 0$, for all $s \neq t$ and all $j, k = 1, \dots, K$; this is not a consequence of (1.2) in general. Without this additional assumption the asymptotic variance of the variance ratio statistics are much more complicated and hard to estimate, involving the selection of a bandwidth parameter. Condition MH1(ii) is satisfied automatically under the martingale hypothesis, which itself is consistent with any kind of nonlinear multivariate ("semi-strong") GARCH process. In Assumption A, we have assumed strict stationarity, whereas this is not required in MH* (although certain sums have to converge in MH3, which would rule out explosive nonstationarity). In MH* we have assumed higher moments depending on the mixing decay rate, whereas for assumption A only four moments are required and no explicit mixing conditions are employed. It should be noted therefore that the conditions A and MH* are non-nested. We further note that under the assumption that returns are i.i.d. (referred to as RW1 in Campbell, Lo, and MacKinlay (1997)), the univariate version of the CLT's below are valid under only second moments, Brockwell and Davis (1991, Theorem 7.2.2), due to the self normalization present in the sample autocorrelations. For similar reasons, condition MH3 may not be strictly necessary in that mildly trending moments may still permit a CLT at the same rate due to the cancellation of numerator by denominator.

We remark that this theory is predicated on the existence of fourth moments, which may be problematic for some financial time series. Provided only the population variance exists, the matrix normalized variance ratio converges in probability to the identity, but may have a non-standard limiting distribution and a slower rate of convergence to it, Phillips and Solo (1992) and Mikosch and Stărică (2000).[12] Even if the population variance does not exist, the sample variance ratio may converge, due to the self-normalization, but one can expect a different scaling law. For example, if the return process is iid with a symmetric stable distribution with parameter $\alpha \in [1, 2]$, then the sample variances scale according to $K^{2/\alpha}$, that is, as $T \to \infty$, $\widehat{\mathcal{VR}}(K) \to K^{(2-\alpha)/\alpha}$ for all K. This is similar asymptotic behaviour to what is found under the bubble process of section 5.2 below when $\alpha = 1$. Wright (2000) has proposed variance ratios based on signs and ranks that are robust to heavy tailed distributions, although require stronger assumptions elsewhere.

1.4.2 Finite/fixed horizon Limiting Distribution Theory

We next present our main results. In this subsection we consider the finite K framework.

Theorem 1.1. *Suppose that either Assumption A or MH* holds. Then, as $T \to \infty$:*

$$\begin{aligned} \sqrt{T}\text{vec}\left(\widehat{\mathcal{VR}}_+(K) - I_d\right) &\implies N\big(0, Q(K)\big) \\ \sqrt{T}\text{vec}\left(\widehat{\mathcal{VR}d}_+(K) - \widehat{Rd}(0)\right) &\implies N\big(0, Qd(K)\big). \end{aligned}$$

It follows that for any vector ω, $\omega^\intercal \text{vec}(\widehat{\mathcal{VR}}_+(K) - I_d)$ is asymptotically normal with mean zero and variance $\omega^\intercal Q(K)\omega/T$. Limiting distributions for smooth functions of the variance ratio matrices can be obtained by the delta method.

[12]For stationary univariate linear processes, the sample autocorrelations can be root-T consistent and asymptotically normal under only second moment assumptions, Brockwell and Davis (1991, Theorem 7.2.2), but this result does not hold for nonlinear processes like GARCH, nor for multivariate linear processes.

For the ordered eigenvalues, we employ a different approach, as they are not smooth functions of the variance ratio matrix under the null hypothesis. Specifically, we use Eaton and Tyler (1991, Theorem 3.2) where it is shown that if the random symmetric matrix $\sqrt{T}(\widehat{\mathcal{VR}}(K) - I_d)$ converges in distribution to a matrix random variable, denoted U, then with $i_d = (1, 1, \dots, 1)^{\intercal}$

$$\sqrt{T}\left(\varphi(\widehat{\mathcal{VR}}(K)) - i_d\right) \Longrightarrow \varphi(U), \tag{1.20}$$

where $\varphi(\widehat{\mathcal{VR}}(K))$ and $\varphi(U)$ are $d \times 1$ vectors of ordered eigenvalues $\widehat{\lambda}_j \in \varphi(\widehat{\mathcal{VR}}(K))$ and $\lambda_j^* \in \varphi(U)$, respectively. It follows for example that

$$\sqrt{T}\left(\widehat{\lambda}_{\max} - 1\right) \Longrightarrow \lambda^*_{\max}(U),$$

whose distribution can be computed by simulation.

1.4.3 Standard Errors and Test Statistics

From the expressions in Theorem 1.1 we can obtain pointwise confidence intervals for scalar functions of the matrices $\widehat{\mathcal{VR}}(K)$ or $\widehat{\mathcal{VR}d}(K) - \widehat{Rd}(0)$ or $\widehat{\mathcal{VR}a}(K)$. Let:

$$\widehat{\Xi}_{jk} = \frac{1}{T} \sum_{t=\max\{j,k\}+1}^{T} \left(X_{t-j} - \overline{X}\right)\left(X_{t-k} - \overline{X}\right)^{\intercal} \otimes \left(X_t - \overline{X}\right)\left(X_t - \overline{X}\right)^{\intercal} \tag{1.21}$$

$$\widehat{Q}(K) = \sum_{j=1}^{K-1}\sum_{k=1}^{K-1} c_{j,K}c_{k,K}\left(\widehat{\Sigma}^{-1/2} \otimes \widehat{\Sigma}^{-1/2}\right)\widehat{\Xi}_{jk}\left(\widehat{\Sigma}^{-1/2} \otimes \widehat{\Sigma}^{-1/2}\right) \tag{1.22}$$

$$\widehat{Q}d(K) = \sum_{j=1}^{K-1}\sum_{k=1}^{K-1} c_{j,K}c_{k,K}\left(\widehat{D}^{-1/2} \otimes \widehat{D}^{-1/2}\right)\widehat{\Xi}_{jk}\left(\widehat{D}^{-1/2} \otimes \widehat{D}^{-1/2}\right),$$

and $\widehat{S}(K) = D_n^+\widehat{Q}(K)D_n^{+\intercal}$ and $\widehat{Sd}(K) = D_n^+\widehat{Qd}(K)D_n^{+\intercal}$, where D_n^+ is the Moore-Penrose pseudoinverse of the duplication matrix, Magnus and Neudecker (1980). Specifically, the asymptotic variance of $\widehat{\mathcal{VR}}d_{ii}(K)$ can be estimated by

$$\widehat{Q}d_{iiii}(K) = \frac{1}{\widehat{\sigma}_{ii}^2}\sum_{j=1}^{K-1}\sum_{k=1}^{K-1} c_{j,K}c_{k,K}\widehat{\Xi}_{jk;iiii} \tag{1.23}$$

$$\widehat{\Xi}_{jk;iiii} = \frac{1}{T}\sum_{t=\max\{j,k\}+1}^{T}\left(X_{it-j} - \overline{X}_i\right)\left(X_{it-k} - \overline{X}_i\right)\left(X_{it} - \overline{X}_i\right)^2$$

$$\widehat{\sigma}_{ii} = \frac{1}{T}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)^2.$$

Note that under the Lo and MacKinlay (1988) condition H4 (i.e. (1.19)) we have $\Xi_{jk} = 0$ for $j \neq k$, so that the asymptotic variance in Theorem 1.1 simplifies, a little. The commonly used asymptotic variance matrix is

$$\widehat{Q}d_{LM}(K) = \sum_{j=1}^{K-1} c_{j,K}^2 \left(\widehat{D}^{-1/2} \otimes \widehat{D}^{-1/2}\right) \widehat{\Xi}_{jj} \left(\widehat{D}^{-1/2} \otimes \widehat{D}^{-1/2}\right), \tag{1.24}$$

whose diagonal elements can be compared with (1.23): they are the same except that $\widehat{\Xi}_{jk;iiii} = 0$ for $j \neq k$. In the iid case, we further have $\Xi_{jj} = \Sigma \otimes \Sigma$ and:

$$Q_{iid}(K) = \sum_{j=1}^{K-1} c_{j,K}^2 I_{d^2} \ ; \ \widehat{Qd}_{iid}(K) = \sum_{j=1}^{K-1} c_{j,K}^2 (\widehat{Rd}\,(0) \otimes \widehat{Rd}\,(0)). \tag{1.25}$$

In the scalar case both these quantities are nuisance parameter free.

As we show in the application, the standard errors derived from (1.22), (1.24), and (1.25) can be quite different. This is also observed in the simulated experiment we conduct later in this chapter. Although there is no necessary ordering, generally speaking the standard errors from $\widehat{Q}(K)$ are larger than the standard errors from $\widehat{Q}_{LM}(K)$, which in turn are larger than the standard errors from the i.i.d special case $\widehat{Q}_{iid}(K)$.

The standard errors for univariate quantities of interest can be obtained from (1.22). Let $\tau_f = f(\mathrm{vec}(\mathcal{VR}_+(K)))$ and $\tau d_f = f(\mathrm{vec}(\mathcal{VR}d_+(K)))$ be scalar parameters of interest, where f is a continuously differentiable function with non-zero gradient, and let $e_f = \nabla f(\mathrm{vec}(\mathcal{VR}_+(K))), ed_f = \nabla f(\mathrm{vec}(\mathcal{VR}d_+(K))) \in \mathbb{R}^{d^2}$ denote the gradients of the functions at the true value. Let $Q_f(K) = e_f^\intercal Q(K) e_f$ and $Qd_f(K) = e_f^\intercal Qd(K) e_f$. Then $\widehat{Q}_f(K) = e_f^\intercal \widehat{Q}(K) e_f$ and $\widehat{Qd}_f(K) = e_f^\intercal \widehat{Qd}(K) e_f$ are consistent asymptotic variance estimators for $\widehat{\tau}_f$ and $\widehat{\tau d}_f$ respectively. For example, define the column vectors: b that is 0 at the $((l-1)(d+1)+1)^{th}$ entries $(l = 1, \ldots, d)$ and 1 otherwise; i is a conformable column vector of ones; c is a column vector that is $(1-d)/(d^2(K-1))$ at $((l-1)(d+1)+1)^{th}$ entries, and is $1/(d^2(K-1))$ at other entries; and $\delta = \mathrm{vech}(I_d)$. Then, specifically, let:

$$\widehat{Q}_{CS}(K) = \frac{1}{d^2(d-1)^2} b^\intercal \widehat{Qd}\,(K)\, b$$

$$\widehat{Q}_{GMV}(K) = d^{-4} i^\intercal \widehat{S}(K) i,$$

$$\widehat{Q}_\pi(K) = c^\intercal \widehat{Qd}(K) c$$

$$\widehat{Q}_{tr}(K) = \delta^\intercal \widehat{S}\,(K)\, \delta = \widehat{Q}_{\det}(K).$$

We next define some test statistics. Let f be any continuously differentiable function with nonzero gradient (for example $CS, \det, GM, tr$, or π), and let

$$Z_f(K) = \sqrt{T}\left(\widehat{Q}_f(K)\right)^{-1/2}\left[f(\text{vec}(\widehat{\mathcal{VR}}_+(K))) - f(\text{vec}(\mathcal{VR}_+(K)))\right] \tag{1.26}$$

$$Zd_f(K) = \sqrt{T}\left(\widehat{Q}d_f(K)\right)^{-1/2}\left[f(\text{vec}(\widehat{\mathcal{VR}d}_+(K))) - f(\text{vec}(\widehat{Rd}(0)))\right] \tag{1.27}$$

$$W_F(K) = T\text{vech}\left(\widehat{\mathcal{VR}}(K) - I\right)^{\intercal}\widehat{S}(K)^{-1}\text{vech}\left(\widehat{\mathcal{VR}}(K) - I\right) \tag{1.28}$$

$$Wd_F(K) = T\text{vech}\left(\widehat{\mathcal{VR}d}(K) - \widehat{Rd}(0)\right)^{\intercal}\widehat{Sd}(K)^{-1}\text{vech}\left(\widehat{\mathcal{VR}d}(K) - \widehat{Rd}(0)\right). \tag{1.29}$$

Corollary 1.1. *Suppose that either Assumption A or MH* holds. Then (for each fixed K) the estimator $\widehat{Q}(K)$ is weakly consistent for $Q(K)$ (likewise, $\widehat{Q}d(K)$ are weakly consistent for $Qd(K)$), i.e., as $T \to \infty$,*

$$\widehat{Q}(K) \xrightarrow{P} Q(K)$$

$$Z_f(K), Zd_f(K) \Longrightarrow N(0,1)$$

$$W_F(K), Wd_F(K) \Longrightarrow \chi^2(d(d+1)/2).$$

In the application we make use of a bias correction method based on asymptotic expansions (under the iid assumption), which may give better performance for long lags. A number of alternative inference methods such as self-normalization, or block bootstrap and subsampling have been suggested to accommodate the more general uncorrelatedness assumption that allows $E\big[\widetilde{X}_t\widetilde{X}_{t-j}^{\intercal} \otimes \widetilde{X}_s\widetilde{X}_{s-k}^{\intercal}\big] \neq 0$ for some $s \neq t$. The readers are directed to Lobato (2001) and Whang and Kim (2003) for description of these methods.

1.4.4 Increasing horizon Limiting Distribution Theory

It has been reported in the literature that inferences based on the asymptotic theory of the variance ratio statistic become unreliable in finite samples when the horizon K is large relative to the sample size T, see Lo and MacKinlay (1989). In view of this practical issue, Richardson and Stock (1989) considered the framework in which $K = K(T)$ and $K/T \to \delta < 1$, and showed that the limiting distribution is a function of Brownian motion. However, Deo and Richardson (2003) pointed out the inconsistency of the univariate variance ratio test under this particular restriction against some important mean reverting alternatives. Consequently, Chen and Deo (2006) studied an alternative setting where K is set to increase slower so that K/T tends to zero. Along with the ergodic martingale difference assumption, they imposed a set of strong conditions on cross-moments (Assumption A3) including the no-leverage condition, and some mixing-type conditions (Assumptions A5 and A6) that imply asymptotic independence of the process.

In this section we investigate the increasing K asymptotics in the multivariate framework. Although a d-dimensional analogue of the conditions assumed in Chen and Deo (2006) can be adopted, we shall consider a different set of conditions including stationarity in Assumption A1 (but with a slightly higher moment condition). This is to be consistent with the previous fixed K theory, and to allow simple derivations under mild assumptions.

Assumption A$'$. *The process $\widetilde{X}_t$ is a stationary ergodic Martingale Difference sequence having finite $4+\delta$ moments, i.e. $E|\widetilde{X}_{it}|^{4+\delta} \leq C < \infty$ for some $\delta > 0$ for all i.*

Assumption T. *The horizon $K \to \infty$ as $T \to \infty$ and $K/T \to 0$.*

Assumption S. *The following double sum is finite: $\sum_{a=-\infty}^{\infty} \sum_{b=-\infty}^{\infty} |\kappa_{pqrl}(a,b,0,0)| < \infty$ for all $p,q,r,l \leq d$, where $\kappa_{pqrl}(t_1,t_2,t_3,t_4)$ is the cumulant of 4th order between $(\widetilde{X}_{pt_1}, \widetilde{X}_{qt_2}, \widetilde{X}_{rt_3}, \widetilde{X}_{lt_4})$.*

Along with stationarity, Assumption S guarantees the existence and positive definiteness of the matrix limits $Q(\infty) = \lim_{K\to\infty} K^{-1}Q(K)$ and $Qd(\infty) = \lim_{K\to\infty} K^{-1}Qd(K)$, both of which will turn out to have simple forms, see the proof of Theorem 2.1 in the appendix. Indeed, summability of the cumulants is a common assumption in the time series literature, see Rosenblatt (1985). The weak condition regulates the dependence structure of the process, and is implied by a mild α-mixing and moment condition (strictly higher than 4 as we shall assume below) as shown in Andrews (1991, Lemma 1), although it is stronger than ergodicity. For example, Assumption MH2 with size of mixing strengthened to $3r/(r-1)$ is sufficient for summability of cumulants under stationarity. Some relevant discussions can be found in the recent paper Shao and Wu (2007), where an alternative sufficient condition is given in terms of the notion of geometric moment contraction (GMC).

We derive the limiting distribution under the stationary ergodic martingale difference assumption. Note that one could alternatively work with mixing (Assumption MH*) or near epoch dependence, for example, to obtain a similar result, but we shall not proceed to this direction in our work here.

Theorem 1.2. *Suppose that Assumptions A$'$, T and S hold. Then, as $T \to \infty$:*

$$\sqrt{\frac{T}{K}}\mathrm{vec}\left(\widehat{\mathcal{VR}}_{+}(K) - I_d\right) \implies N\left(0, \frac{4}{3}J\right)$$
$$\sqrt{\frac{T}{K}}\mathrm{vec}\left(\widehat{\mathcal{VR}d}_{+}(K) - I_d\right) \implies N\left(0, \frac{4}{3}Jd\right),$$

and

$$\begin{aligned} Z_f(K), Zd_f(K) &\Longrightarrow N(0,1) \\ W_F(K), Wd_F(K) &\Longrightarrow \chi^2(d(d+1)/2), \end{aligned}$$

where J is the idendty matrix of dimension $d^2 \times d^2$, and $Jd = (D^{-1/2} \otimes D^{-1/2})(\Sigma \otimes \Sigma)(D^{-1/2} \otimes D^{-1/2})$ is the matrix whose diagonal entries are one.

This says that the inference methods we apply in the finite K case can be carried over to the increasing K case, at least where K is not too large relative to the sample size. Chen and Deo (2006) has some discussions about the finite sample performance with respect to different K/T ratios.

1.5 Alternative Hypotheses

There are many plausible alternative hypotheses to the null hypothesis (1.2), and it is not possible in general to have power against all such departures. We can understand a little bit better the type of alternatives against which the variance ratio has power by looking at equation (1.12). We have $\mathcal{VR}(K) = I$ if and only if

$$\sum_{j=1}^{K-1} c_{j,K} \left(\mathcal{R}(j) + \mathcal{R}(j)^{\intercal}\right) = \mathbf{0}.$$

This says that the test will have power against alternatives for which the Bartlett weighted autocorrelations do not sum identically to zero. This seems like a reasonable class of alternative, because if the autocorrelations change sign enough that they cancel out, this seems like a not very propitious setting to make excess returns from a trading strategy that treats these autocorrelations as signals. One wants not just departures from zero but some kind of reliable direction of dependence on which to bet. By contrast, the Box-Pierce statistic will also pick up highly oscillatory variation in the autocorrelations, which one might prefer to exclude from consideration.

We look in detail at several alternative models in this section. In general they yield a prediction of the form

$$\Sigma_T(K) = K\Sigma + \Delta(K,T), \tag{1.30}$$

where $\Delta(K,T)$ is a symmetric matrix such that $\Sigma_T(K) > 0$.

1.5.1 Local Alternatives

We first extend the arguments presented by Faust (1992) to the multivariate case and show that a trace test will be optimal against a certain class of alternatives. The type of mean reversion that the test is best at detecting will be shown to be a special case of vector autoregressive processes of order $K-1$. The main idea is to find a statistic that is *asymptotically equivalent* to the likelihood ratio statistic, since in such a case the test based on that statistic will possess the same local large-sample optimality properties of LR tests, see Engle (1984). Below we show that the statistic based on $\mathrm{tr}(\widehat{\mathcal{VR}}(K))$ (defined formally below) is optimal (under normality) for testing the null hypothesis of no predictability/serial correlation, against the alternative hypothesis that *each marginal* process $\{X_{jt}\}_t$, $j = 1, \ldots, d$ belongs to what is called the $\phi-$*best* class proposed by Faust (1992). The $\phi-$best class is a particular class of $AR(K-1)$ models, and is defined as the set of those having AR polynomials $\rho_q(L)$ that satisfy

$$\rho_q(z)\rho_q(z^{-1}) = \alpha\big(1 + q\phi(z)\phi(z^{-1})\big) \tag{1.31}$$

for some constants q and $\alpha > 0$, and z inside the unit circle; the coefficients for the moving average filter $\phi(L)$ are $\phi_j = +1$ for all $j = 0, \ldots, K-1$. From the definition we see that under the alternative hypothesis, $\{X_t\}$ essentially belongs to a (particular) class of *vector* autoregressive process $VAR(K-1)$. We note that when $q = 0$ the process is a white noise. Denote by $\mathbf{X}$ the $T \times d$ matrix of sample observations. Then formally, the null and alternative hypotheses can be written as

$$\begin{aligned} &H_0 : \mathbf{X} \sim \mathcal{N}_d^T\left(i\mu^\intercal, I_T \otimes \Sigma\right) && \text{[Uncorrelatedness]} \\ &H_1 : \mathbf{X} \sim \mathcal{N}_d^T\left(i\mu^\intercal, \Sigma_{q*} \otimes \Sigma\right) && \text{[`}\phi-\text{best' temporal dependence]} \end{aligned}$$

where Σ_{q^*} refers to the variance-covariance matrix of the $\phi-$ best class process with the index of the process $q = q^* > 0$. The notation $\mathcal{N}_d^T$ stands for a matrix normal variable; each matrix (separated by the Kronecker product) in the variance represents the contribution from cross-sectional and temporal sides, respectively. So essentially, this is a one-sided test of the index q being zero *versus* q being a strictly positive constant. Examination of the local large-sample optimality is done by letting the index $q^* = q^*(T) = \delta/\sqrt{T}$ in the alternatives, where δ determines the direction to which the test departs from the null hypothesis.

Theorem 1.3. *Suppose that the data are normally distributed. Then, the trace test is locally most powerful invariant against alternatives in the $\phi-$best class of the form $q_T^* = \delta/\sqrt{T}$.*

It may be possible to characterize the class of alternatives against which other tests, such as the determinant test, are optimal, but we leave this for future research.

The trace test, while optimal against the specific class above, may have zero power against some alternatives, as we next discuss.

Suppose that $\Delta(K,T) = \Delta(K)/\sqrt{T}$, then

$$\begin{aligned}
\sqrt{T}\,(\mathcal{VR}(K) - I) &= \frac{1}{K}\Sigma^{-1/2}\Delta(K)\Sigma^{-1/2}; \\
\sqrt{T}\,(\mathcal{VR}d(K) - Rd(0)) &= \frac{1}{K}D^{-1/2}\Delta(K)D^{-1/2}.
\end{aligned}$$

Provided $\Delta(K)$ is strictly definite, some tests based on these matrices will have positive power against this alternative. On the other hand, in some cases, the power may be zero. Specifically, suppose we take the trace test applied to the diagonally normalized variance ratio matrix, i.e., compare $\mathrm{tr}(\widehat{\mathcal{VR}}d(K)) - d$ (c.f. Castura et al. (2010)) with the critical values from its normal limit given above, then if $\Delta(K)$ is of the form

$$\Delta_{ij}(K) = \begin{cases} \delta(K) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

for some nonzero $\delta(K)$, then this particular test will have zero power.

1.5.2 Multivariate Fads Model

We consider an alternative to the efficient market hypothesis (1.2), which allows for temporary mispricing through fads but assures that the rational price dominates in the long run. Consider the multivariate fads model for log prices:

$$p_t^* = \mu + p_{t-1}^* + \varepsilon_t \tag{1.32}$$

$$p_t = p_t^* + \eta_t, \tag{1.33}$$

where ε_t is iid with mean zero and variance matrix Ω_ε, while η_t is a stationary weakly dependent process with unconditional variance matrix Ω_η, and the two processes are mutually independent. It follows that the observed return satisfies

$$X_t = p_t - p_{t-1} = \mu + \varepsilon_t + \eta_t - \eta_{t-1}. \tag{1.34}$$

This is a multivariate generalization of the scalar Muth (1960) model, which was also adopted in Poterba and Summers (1988). It allows actual prices p to deviate from fundamental prices p^* but only in the short run through the fad process η_t. This process is a plausible alternative to the efficient markets hypothesis. If η_t were i.i.d., then X_t would be (to second order) an MA(1) process, which is a structure implied by a number of market microstructure issues (Hasbrouck (2007)). In this case,

$$\mathcal{VR}(K) = I + (1 - \frac{1}{K})(R(1) + R(1)^\intercal) = I - 2(1 - \frac{1}{K})\left(\Omega_\varepsilon + 2\Omega_\eta\right)^{-1/2}\Omega_\eta\left(\Omega_\varepsilon + 2\Omega_\eta\right)^{-1/2},$$

and likewise for $\mathcal{VR}d(K)$. In general, however, η_t might have any type of weak dependence structure.

We next derive a restriction on the long run variance ratio statistic that reflects the presence of fads. We do not restrict the fads process, and so can only obtain long run implications.

Theorem 1.4. *Suppose that the multivariate fads model (1.32)-(1.33) holds and suppose that* $\mathrm{cov}(\eta_{t+j}, \eta_t) \to 0$ *as* $j \to \infty$. *Then,* $\mathcal{VR}(\infty) = \lim_{K\to\infty} \mathcal{VR}(K) = I + \sum_{j=1}^{\infty}(R(j) + R(j)^{\intercal})$ *exists. Further suppose that* $\Omega_\eta(1) \equiv \mathrm{var}(\eta_t - \eta_{t-1}) > 0$. *Then,*

$$\mathcal{VR}(\infty) < I_d$$

in the matrix partial order sense. Likewise, $\mathcal{VR}d(\infty) = \lim_{K\to\infty} \mathcal{VR}d(K)$ *exists, and*

$$\mathcal{VR}d(\infty) < Rd(0).$$

This result generalizes the existing results for the scalar fads process, which amount to $\mathcal{VR}d_{ii}(\infty) \leq Rd_{ii}(0)$ for $i = 1, \ldots, d$. In Theorem 1.4, we obtain stronger constraints on the off diagonal elements of $\mathcal{VR}d(\infty)$ and $\mathcal{VR}(\infty)$. Note that we also obtain $GMV(K) \to GMV(\infty) > 1/d$ as a corollary.

We consider what happens to the long horizon sample variance ratio statistic under the fads model. We will consider the case where $K \to \infty$ as $T \to \infty$ such that $K/T \to 0$ (in contrast with the framework of Richardson and Stock (1989)). The consistency follows from the theory for the long run variance ratio, Parzen (1957), Andrews (1991), and Liu and Wu (2010). We adopt the framework of Liu and Wu (2010) and suppose that

$$X_t = \Psi\left(\ldots, e_{t-1}, e_t\right),$$

where e_t are i.i.d random vectors of length $p \geq d$, and $\Psi : \mathbb{R}^p \times \mathbb{R}^p \times \cdots \to \mathbb{R}^d$. This includes a wide range of linear and nonlinear processes for η_t, ε_t. Then define

$$\delta_t = E\left[||(\Psi\left(\ldots, e_0, \ldots, e_{t-1}, e_t\right) - \Psi\left(\ldots, e_0', \ldots, e_{t-1}, e_t\right))||\right],$$

where e_t' is an i.i.d. copy of e_t and $||.||$ denotes the Euclidean norm.

Assumption B. *The vector process* X_t *is stationary with finite fourth moments and weakly dependent in the sense that* $\sum_{t=1}^{\infty} \delta_t < \infty$.

Theorem 1.5. *Suppose that the multivariate fads model (1.32)-(1.33) holds along with*

Assumption B, and suppose that $K \to \infty$ as $T \to \infty$ such that $K/T \to 0$. Then,

$$\widehat{\mathcal{VR}}(K) \xrightarrow{P} \mathcal{VR}(\infty).$$

Likewise, $\widehat{\mathcal{VR}d}(K)$ consistently estimates $\mathcal{VR}d(\infty)$. More generally, we could obtain the limiting distribution of $\widehat{\mathcal{VR}}(K) - \mathcal{VR}(K)$ under either fixed K or K increasing asymptotics applying the methods of Liu and Wu (2010), but the limiting variance in either case is going to be very complicated.

1.5.3 Bubble Process

Several authors argue that the frequently observed excessive volatility in stock prices may be attributed to the presence of speculative bubbles. Blanchard and Watson (1982) and Flood and Hodrick (1986), inter alia, demonstrate in a theoretical framework that bubble components potentially generate excessive volatility. There is some debate about whether these constitute rational adjustment to fundamental pricing rules or arise from more behavioural reasons. Recently, Phillips and Yu (2011), and Phillips, Shi, and Yu (2012) have considered the following class of "bubble processes" for (log) prices p_t

$$\begin{aligned} p_t &= \mu + p_{t-1} 1\left(t < \tau_e\right) + \delta_T 1\left(\tau_e \le t \le \tau_f\right) p_{t-1} \\ &\quad + \left(\sum_{s=\tau_f+1}^{t} \varepsilon_s + p_{\tau_f}^* \right) 1\left(t > \tau_f\right) + \varepsilon_t 1\left(t \le \tau_f\right), \end{aligned} \tag{1.35}$$

where $p_{\tau_f}^*$ represents the restarting price after the bubble collapses at time τ_f, and $\delta_T = 1 + c/T^{\alpha}$ for $\alpha \in (0,1)$ and $c > 0$. The process is consistent with the efficient markets hypothesis during $[1, \tau_e]$ and $[\tau_f, T]$ but has an explosive "irrational" moment in the middle. They propose econometric techniques to test for the presence of a bubble and indeed multiple bubbles. One can imagine this model also holding for a vector of asset prices caught up in the same bubble, so that ε_t is a vector of shocks, the indicator function is applied coordinatewise, and the coefficient δ_T is replaced by a diagonal matrix.

In the appendix we show that in the univariate bubble process with nontrivial bubble epoch (i.e., $(\tau_f - \tau_e)/T \to \tau_0 > 0$), that, as $T \to \infty$

$$\widehat{\mathcal{VR}}(K) \xrightarrow{P} K \tag{1.36}$$

for all K, so that the variance ratio statistic is greater than one for all K and gets larger with horizon. Essentially, the bubble period dominates all the sample statistics, and all return autocorrelations converge to one inside the bubble period, thereby making the ratio equal to the maximum it can achieve. In the multivariate case, Magdalinos (2014) has shown that in some special cases, $\lambda_{\max}(\widehat{\mathcal{VR}}(K)) \xrightarrow{P} K$. However, the multivariate case is more complicated because other eigenvalues may not behave in the same way.

In practice, rolling window versions of the variance ratio statistics can detect the bubble period in a similar way to the Phillips, Shi and Yu (2012) statistics (although they are not explicitly designed for this purpose and are not optimal for it). Our point here is just that these two different alternative models generate opposite predictions with regard to the variance ratio. We will check this empirically below.

1.5.4 Time Varying Expected Return

We briefly consider a simple statistical model for time varying expected return. This model could be consistent with rational pricing where the risk premium evolves slowly over time and has small variation relative to the shocks to risk adjusted returns. Specifically, suppose that observed returns are composed of a slowly varying risk premium μ_t and an iid shock ε_t, i.e.,

$$X_t = \mu + \mu_t + \varepsilon_t, \tag{1.37}$$

$$\mu_t = \mu_{t-1} + \eta_t, \tag{1.38}$$

where $\mu_0 = (0, 0, \ldots, 0)^\intercal$ and η_t is an iid mean zero shock that is "small" relative to ε_t. In this case observed returns are nonstationary so we must index populations by T. This specification is similar to that of equation 7.1.30 of CLM. We establish the following result.

Theorem 1.6. *Suppose that the model (1.37)-(1.38) holds with η_t iid mean zero with $E\eta_t\eta_t^\intercal = \Sigma_\eta/T > 0$ and $\varepsilon_t$iid mean zero with $E\varepsilon_t\varepsilon_t^\intercal = \Sigma_\varepsilon > 0$. Then*

$$\lim_{K\to\infty} K^{-1} \lim_{T\to\infty} \mathcal{VR}_T(K) \to \left(\Sigma_\eta \frac{1}{2} + \Sigma_\varepsilon\right)^{-1/2} \frac{1}{2}\Sigma_\eta \left(\Sigma_\eta \frac{1}{2} + \Sigma_\varepsilon\right)^{-1/2} < I_d.$$

This model gives a similar prediction to the bubble model, except it says that all eigenvalues should grow linearly with the horizon with a slope less than one.

In the working paper version of this chapter we consider several alternative approaches to capturing time varying expected returns including nonparametric mean model and linear factor models.

1.5.5 Locally Stationary Alternatives

Suppose that $X_t = X_{t,T}$ can be approximated by a family of locally stationary processes $\{X_t(u),\, u \in [0, 1]\}$, Dahlhaus (1997). For example, suppose that $X_t = \varepsilon_t + \Theta(t/T)\varepsilon_{t-1}$, where $\Theta(\cdot)$ is a matrix of smooth functions and ε_t is iid. This allows for zones of departure from the null hypothesis, say for $u \in U$, where U is a subinterval of $[0,1]$, e.g., $\Theta(u) \neq 0$ for $u \in U$. For example, during recessions the dependence structure may change and depart from efficient markets, but return to efficiency during normal times. This is consistent with the Adaptive Markets Hypothesis of Lo (2004, 2005) whereby the amount of inefficiency can change over time depending on "the number of competitors in the market, the magnitude of profit opportunities available, and the adaptability of the market participants".

Let $\widetilde{X}_t(u) = X_t(u) - EX_t(u)$ and:

$$\Sigma(u) = \mathrm{var}(X_t(u)) = E(\widetilde{X}_t(u)\widetilde{X}_t^{\intercal}(u))$$
$$D(u) = \mathrm{diag}\left\{E(\widetilde{X}_{1t}^2(u)), \ldots, E(\widetilde{X}_{dt}^2(u))\right\}$$
$$\Gamma(j;u) = E(\widetilde{X}_t(u)\widetilde{X}_{t-j}^{\intercal}(u)).$$

The sample autocovariances converge, under some conditions, to the integrals of the autocovariances, e.g., $\widehat{\Gamma}(j) \to \int_0^1 \Gamma(j;u)du$. Then, define

$$\overline{R}(j) = \left(\int_0^1 \Sigma(u)du\right)^{-1/2} \int_0^1 \Gamma(j;u)du \left(\int_0^1 \Sigma(u)du\right)^{-1/2}.$$

It follows that under local stationarity

$$\widehat{\mathcal{VR}}(K) \xrightarrow{P} \overline{\mathcal{VR}}(K) = I + \sum_{j=1}^{K-1}\left(1 - \frac{j}{K}\right)(\overline{R}(j) + \overline{R}(j)^{\intercal}).$$

The test will have power against some alternatives where $\Gamma_u(j) \neq 0$ for $u \in U$ and $\Gamma_u(j) = 0$ for $u \in U^c$. The test will not detect alternatives where $\overline{\mathcal{VR}}(K) = I$ but $VR(K;u) = I + \sum_{j=1}^{K-1}(1 - \frac{j}{K})(R(j;u) + R(j;u)^{\intercal}) \neq 0$, where $R(j;u) = \Sigma(u)^{-1/2}\Gamma(j;u)\Sigma(u)^{-1/2}$.

1.5.6 Nonlinear Processes

In general, the class of statistics we consider will not have power against all nonlinear alternatives. In that case, one may work with nonlinear transformations $Y_t = \tau(X_t)$ such as the quantile hit process, Han et al. (2014), and then calculate the "variance ratio" equivalent through (1.12)-(1.14). Wright (2000) has proposed variance ratios based on signs and ranks that have similar objectives.

1.6 Large Dimensional Data

We briefly consider some issues that arise when the dimensions d are large. In this case, the covariance matrices Σ and $\Sigma(K)$ may be ill conditioned, and so forming the ratio

(1.11) may not be practically feasible or theoretically valid; likewise for any functions derived thereof such as the smallest eigenvalues. The diagonal variance ratio matrix and simple univariate quantities derived from it like $CS(K)$ may fare better in this situation, since the marginal variances should be bounded away from zero. We remark that Castura, Litzenberger, Gorelick, and Dwivedi (2010) report the average variance ratio of the Russell 1000 and Russell 2000 stocks, which amounts to $\sum_{i=1}^{d} \widehat{\mathcal{VR}d}_{ii}(K)/d$. They do not report standard errors for this quantity, perhaps on the grounds that d is large (since $d = 3000$). However, when the individual stocks are contemporaneously correlated, which they typically are[13], the averaging will not reduce the order of magnitude of the standard error. Specifically, under the iid assumption, the correlation between $\widehat{\mathcal{VR}d}_{ii}(K)$ and $\widehat{\mathcal{VR}d}_{jj}(K)$ will be proportional to ρ_{ij}^2, where ρ_{ij} is the contemporaneous correlation between the returns on stock i and stock j. We show below how to calculate the standard errors for $\sum_{i=1}^{d} \widehat{\mathcal{VR}d}_{ii}(K)/d$ in the large d, T case. However, for nonlinear functions of $\mathcal{VR}d(K)$ such as its eigenvalues, or for quantities derived from $\mathcal{VR}(K)$, the large d theory is more complicated.

We present a simple result for the average trace statistic in the case where d grows but at a rate slower than T. We suppose that Assumption A$'$ holds for the d-dimensional vector process $\widetilde{X}_t$, and impose Assumption Sd below to ensure that the limiting variance is well-defined.

Assumption Td. *The dimension $d = d(T) \to \infty$ in such a way that $d/T \to 0$ as $T \to \infty$.*

Assumption Sd. *The limit of the quadruple sum $qd(\infty)^* \in (0, \infty)$ exists, where*

$$qd(\infty)^* := \lim_{d\to\infty} \frac{1}{d^2} \sum_{i=1}^{d} \sum_{r=1}^{d} \left(\sum_{j=1}^{K-1} \sum_{k=1}^{K-1} \frac{c_{j,K} c_{k,K}}{\sigma_{ii}\sigma_{rr}} E\left[\widetilde{X}_{it} \widetilde{X}_{rt} \widetilde{X}_{i,t-j} \widetilde{X}_{r,t-k} \right] \right), \tag{1.39}$$

where σ_{ii} are the diagonal elements of Σ.

Under these conditions, we can derive the following asymptotic normality result:

Theorem 1.7. *Suppose that Assumptions A$'$, S, Td, and Sd hold. Then:*

$$Zd_{tr}(K) \Longrightarrow N(0, 1).$$

We remark that the cross-sectional standard deviation of the individual variance ratios, a quantity that is often reported along with the average variance ratio, see for ex-

[13] Although for very high frequency data, the correlation maybe quite small, Sheppard (2013).

ample CLM Table 2.7, is not necessarily related in any simple way to the true asymptotic standard deviation of the estimator that we report here.

An alternative strategy in the large d case may be to calculate scalar ratios from the matrix scaling law $\Sigma(K) = K\Sigma$. Specifically, we may look at quantities like $\lambda_{\max}(K)$ over $K\lambda_{\max}(1)$ whose properties may follow from generalizations of results in Jin, Wang, Bai, Krishnan, and Harding (2014). However, when d is comparable with T, one must use some sparsity structure or shrinkage method to obtain reasonable performance for complicated nonlinear functions of the covariance matrices. Johnstone and Onatski (2015) develop a comprehensive theory for multivariate testing in large dimensional situations.

1.7 Simulation Study

We perform a small simulation study to assess the reliability of our multivariate variance ratio test statistics (the earlier version of this chapter contains additional results not reported here for brevity). In particular we examine two multivariate variance ratio tests: the trace ($Z_{tr}(K)$) and the determinant ($Z_{\det}(K)$) tests.

We first simulate empirical size of nominal 5% multivariate variance ratio tests based on $Z_{tr}(K)$ and $Z_{\det}(K)$ statistics for the null hypothesis $H_0 : X_t = (X_{1,t}, X_{2,t})^{\intercal}$ is m.d.s. specified by the following bivariate constant conditional correlation (CCC)-GARCH (1,1) model:

$$X_{1,t} = \sqrt{h_{1,t}}\varepsilon_{1,t},\ X_{2,t} = \sqrt{h_{2,t}}\varepsilon_{2,t}$$

$$h_{1,t} = 0.2 + 0.05X_{1,t-1}^2 + 0.9h_{1,t-1}$$

$$h_{2,t} = 0.1 + 0.08X_{2,t-1}^2 + 0.9h_{2,t-1}$$

$$\begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \rho = 0.5.$$

Based on 10000 replications, we have the following results.

Table 1.1: Empirical size of nominal 5% multivariate variance ratio tests [using $Z_{tr}(K)$ and $Z_{\det}(K)$ statistics]

		Size of 5 percent test	
Sample size	K	$Z_{tr}(K)$	$Z_{\det}(K)$
1024	2	0.0488	0.0481
1024	4	0.0478	0.0455
1024	8	0.0467	0.0437
1024	16	0.0507	0.0422

Table 1.1 shows that the empirical sizes of variance ratio tests using $Z_{tr}(K)$ and $Z_{\det}(K)$ statistics are all close to the nominal value 5%. We then examine the power of multivariate variance ratio tests based on $Z_{tr}^{(iid)}(K)$ and $Z_{\det}^{(iid)}(K)$ statistics, at a 5% nominal level, against the alternative hypotheses H_1 : bivariate fads model for log prices, specified as (1.32) and (1.33) with $\mu = 0$ and $\eta_t = \beta\eta_{t-1} + \xi_t$, where $\varepsilon_t \sim i.i.d.N(0, \Omega_\varepsilon)$, $\xi_t \sim i.i.d.N(0, I_d)$, ε_t and η_t are mutually independent, $\beta = \begin{bmatrix} 0.95 & 0.02 \\ 0.05 & 0.9 \end{bmatrix}$. We consider three cases: $\Omega_\varepsilon = 2I_d$, $\Omega_\varepsilon = I_d$ and $\Omega_\varepsilon = \frac{1}{2}I_d$, so that the conditional variability of the random walk relative to the stationary component is two, one and one-half, respectively. We consider $Z_{tr}^{(iid)}(K)$ and $Z_{\det}^{(iid)}(K)$ statistics which are similarly defined as $Z_{tr}(K)$ and $Z_{\det}(K)$ but using $\widehat{Q}_{iid}(K)$. Based on 10000 replications, we have the following results.

Table 1.2: Power of multivariate variance ratio tests at a 5% nominal level [using $Z_{tr}^{(iid)}(K)$ and $Z_{\det}^{(iid)}(K)$ statistics]

		$\Omega_\varepsilon = \frac{1}{2}I_d$		$\Omega_\varepsilon = I_d$		$\Omega_\varepsilon = 2I_d$	
Sample size	K	$Z_{tr}^{(iid)}(K)$	$Z_{\det}^{(iid)}(K)$	$Z_{tr}^{(iid)}(K)$	$Z_{\det}^{(iid)}(K)$	$Z_{tr}^{(iid)}(K)$	$Z_{\det}^{(iid)}(K)$
1024	2	0.2021	0.1971	0.1357	0.1324	0.0844	0.0813
1024	4	0.3933	0.3806	0.2399	0.2273	0.1317	0.1216
1024	8	0.6334	0.6183	0.3932	0.3658	0.1980	0.1728
1024	16	0.8229	0.8009	0.5331	0.4716	0.2653	0.2061

Table 1.2 shows that the power of the tests increases with K. In addition, as the conditional variability of the random walk relative to the stationary component decreases, the power of tests increases, for example, when $\Omega_\varepsilon = \frac{1}{2}I_d$ and $K = 16$, the power of tests is very high which goes beyond 80%. Furthermore, we found the tests based on $Z_{tr}^{(iid)}(K)$ statistics are more powerful than those based on $Z_{\det}^{(iid)}(K)$ statistics across all cases.

1.8 Application

We apply our methodology to U.S. stock return data. In particular, we use weekly size-sorted equal-weighted portfolio returns from the Center for Research in Security Prices (CRSP) from 06/07/1962 to 27/12/2013.[14] Essentially we are using the extension of the same data that were used in Lo and MacKinlay (1988) and Campbell, Lo and Mackinlay (1997), which allows us to make comparison with their results, and to extend it to the more recent period. In the following parts, we first test the linear predictability for

[14]The data are obtained from Kenneth French's Data Library. It was created by CMPT_ME_RETS using the 2013/12 CRSP database. We compute weekly returns of portfolios by linearly adding up Monday to Friday's daily returns.

size-sorted CRSP portfolio returns at short to medium horizon; then we investigate the long-run behavior of variance ratio statistics.

1.8.1 Short to Medium Horizon

(1) Evidence on Linear Predictability

Consider weekly returns for three size-sorted CRSP portfolios $X_t = (X_{1t}, X_{2t}, X_{3t})^\intercal$, where X_{1t} is for the portfolio of small-size firms (first quintile), X_{2t} is for the portfolio of medium-size firms (third quintile), and X_{3t} is for the portfolio of large-size firms (fifth quintile). $\widehat{\mathcal{VR}d}_+(K)$ and $\widehat{Rd}\,(0)$ of X_t can be estimated based on the method in Section 1.3.

We first test the absence of serial correlation in each of these three weekly size-sorted portfolio returns. As we stated above, the diagonal elements of $\mathcal{VR}d_+(K)$ correspond to the univariate variance ratio statistics, for example, $[\mathcal{VR}d_+(K)]_{11}$ is the variance ratio of small-size portfolio returns. For each $i = 1, 2, 3$, we test the hypotheses of H_0 : $[\mathcal{VR}d_+(K)]_{ii} = 1$ against H_1 : $[\mathcal{VR}d_+(K)]_{ii} \neq 1$. To compare with the results reported in Campbell, Lo and Mackinlay (1997, page 71, Table 2.6), we report $\left[\widehat{\mathcal{VR}d}_+(K)\right]_{ii}$ at $K = 2, 4, 8, 16$ and the corresponding $Zd(K)$, $Zd_{LM}(K)$ and $Zd_{iid}(K)$ statistics[15] in three subsamples: 62:07:06-78:09:29 (848 weeks), 78:10:06-94:12:23 (847 weeks) and 94:12:30-13:12:27 (992 weeks). Subsamples are considered to see if there has been changes in variance ratio over time. Table 1.3A, Table 1.3B and Table 1.3C report the results for small-size portfolio, medium-size portfolio and large-size portfolio, respectively.

Table 1.3A: Variance ratios for weekly small-size portfolio returns

		Lags			
Sample period	# of obs	$K = 2$	$K = 4$	$K = 8$	$K = 16$
62:07:06—78:09:29	848	**1.43**	**1.93**	**2.46**	**2.77**
		(8.82)*	(8.49)*	(7.00)*	(5.59)*
		(8.82)*	(10.81)*	(11.00)*	(9.33)*
		(12.46)*	(14.47)*	(14.39)*	(11.70)*
78:10:06—94:12:23	847	**1.43**	**1.98**	**2.65**	**3.19**
		(6.20)*	(7.07)*	(7.37)*	(6.48)*
		(6.20)*	(8.62)*	(10.69)*	(10.70)*
		(12.52)*	(15.25)*	(16.26)*	(14.45)*
94:12:30—13:12:27	992	**1.21**	**1.47**	**1.7**	**1.82**
		(3.30)*	(3.58)*	(3.35)*	(2.50)*
		(3.30)*	(4.13)*	(4.15)*	(3.44)*
		(6.59)*	(7.91)*	(7.43)*	(5.82)*

[15]For testing $[\mathcal{VR}d_+(K)]_{ii} = 1$, the $Zd(K)$, $Zd_{LM}(K)$ and $Zd_{iid}(K)$ statistics are calculated by setting e_f as a column vector that is 1 at the $d(i-1)+i$ entry and 0 otherwise.

Table 1.3B: Variance ratios for weekly medium-size portfolio returns

		Lags			
Sample period	# of obs	$K=2$	$K=4$	$K=8$	$K=16$
62:07:06—78:09:29	848	**1.25**	**1.54**	**1.79**	**1.91**
		$(5.41)^*$	$(5.55)^*$	$(4.35)^*$	$(3.22)^*$
		$(5.41)^*$	$(6.41)^*$	$(5.93)^*$	$(4.69)^*$
		$(7.37)^*$	$(8.42)^*$	$(7.78)^*$	$(6.05)^*$
78:10:06—94:12:23	847	**1.20**	**1.37**	**1.54**	**1.56**
		$(3.29)^*$	$(3.35)^*$	$(3.18)^*$	$(2.14)^*$
		$(3.29)^*$	$(3.72)^*$	$(3.90)^*$	$(2.93)^*$
		$(5.73)^*$	$(5.80)^*$	$(5.36)^*$	$(3.74)^*$
94:12:30—13:12:27	992	**0.99**	**1.05**	**1.02**	**0.89**
		(-0.02)	(0.38)	(0.10)	(-0.38)
		(-0.02)	(0.43)	(0.11)	(-0.48)
		(-0.04)	(0.78)	(0.20)	(-0.78)

Table 1.3C: Variance ratios for weekly large-size portfolio returns

		Lags			
Sample period	# of obs	$K=2$	$K=4$	$K=8$	$K=16$
62:07:06—78:09:29	848	**1.05**	**1.15**	**1.21**	**1.19**
		(1.05)	(1.64)	(1.23)	(0.68)
		(1.05)	(1.54)	(1.32)	(0.84)
		(1.59)	$(2.33)^*$	$(2.06)^*$	(1.29)
78:10:06—94:12:23	847	**1.03**	**1.06**	**1.08**	**1.01**
		(0.63)	(0.61)	(0.54)	(0.03)
		(0.63)	(0.65)	(0.59)	(0.04)
		(0.95)	(0.91)	(0.75)	(0.04)
94:12:30—13:12:27	992	**0.93**	**0.94**	**0.89**	**0.81**
		(-0.99)	(-0.46)	(-0.53)	(-0.62)
		(-0.99)	(-0.52)	(-0.61)	(-0.77)
		$(-2.05)^*$	(-1.01)	(-1.14)	(-1.35)

$\left[\widehat{\mathcal{VR}d}_+(K)\right]_{ii}$ for $i=1,2,3$ are reported in the main rows. Test statistics ($Zd(K)$, $Zd_{LM}(K)$ and $Zd_{iid}(K)$) in parentheses marked with asterisks indicate that the variance ratios are statistically different from one at 5% level of significance.

The results for the earlier sample periods are broadly similar to those in Campbell, Lo and Mackinlay (1997, page 71, Table 2.6) who compared the period 1962-1978 with the period 1978-1994 as well as the combined period 1962-1994. The variance ratios are greater than one and deviate further from one as the horizon lengthens. The departure from the random walk model is strongly statistically significant for the small and medium sized firms, but not so for the larger firms.

When we turn to the later period 1994-2013 we see that the variance ratios all reduce in magnitude. For the smallest stocks the statistics are still significantly greater than one and increase with horizon. However, they are much closer to one at all horizons and the statistical significance of the departures is substantially reduced. For medium sized firms, the variance ratios are reduced. They are in some cases below one and also no longer increasing with horizon. They are insignificantly different from one. For the largest firms, the ratios are all below one but are statistically inseparable from this value. One interpretation of these results is that the stock market (at the level of these portfolios) has become closer to the efficient benchmark. This is consistent with the evidence presented in Castura, Litzenberger, Gorelick, and Dwivedi (2010) for high frequency stock returns. The biggest improvements seem to come in the most recent period, especially for the small stocks.

The test statistics change quite a lot depending on which covariance matrix $\widehat{Q}(K)$, $\widehat{Q}_{LM}(K)$ or $\widehat{Q}_{iid}(K)$ one uses, and in some cases this could affect one's conclusions, for instance, for large-size portfolio, test statistics based on $\widehat{Q}_{iid}(K)$ in some periods are statistically significant. Our sample size is relatively large, and for smaller samples, the differences could matter a lot more.

We test whether the variance ratio has "improved" significantly from one period (A) to the next (B). For this purpose we consider statistics of the form

$$\tau_{AB} = f\left(\widehat{\mathcal{VR}d}_{+}^{A}(K)\right) - f\left(\widehat{Rd}^{A}(0)\right) - f\left(\widehat{\mathcal{VR}d}_{+}^{B}(K)\right) + f\left(\widehat{Rd}^{B}(0)\right), \qquad (1.40)$$

where $\widehat{\mathcal{VR}d}_{+}^{j}(K)$ and $\widehat{Rd}^{j}(0)$ denote the variance ratio statistic and the sample correlation matrix computed in period $j = A, B$, while f is some scalar valued smooth function such as the trace or determinant. Under the martingale null hypothesis (and assuming each subsample is large), the two subsample variance ratio statistics are asymptotically independent and the asymptotic variance of $\sqrt{T}vec(\tau_{AB})$ is just the sum of the subperiod covariance matrices $Qd_f^A(K) + Qd_f^B(K)$. For example, we may consider the single element of statistic $[\widehat{\mathcal{VR}d}_{+}^{A}(K)]_{ii} - [\widehat{\mathcal{VR}d}_{+}^{B}(K)]_{ii}$ and compare it with the square root of the sum of the square of the associated standard errors to obtain a test of the hypothesis that the efficiency has not improved across subperiods. For example, in Table 1.3A, the change of the variance ratio for small stocks of 1.43 in the period 78:10:06-94:12:23 to 1.21 during 94:12:30-13:12:27 is statistically significant according to this calculation.

We have carried out this calculation using the Friday to Friday weekly returns as the base series, but we have also done it for other days of the week and for the two parameter statistic. Qualitatively the results are similar. Results are available upon request.

(2) Lead Lag Relationships

In addition to the autocorrelation for each asset, the predictability can also come from the cross-autocorrelation (lead-lag relationship) between the assets. As we stated earlier, the off-diagonal elements of $\mathcal{VR}d_+(K) - Rd(0)$ provide information about the cumulative cross-dynamics between the assets. We test the hypothesis of $[\mathcal{VR}d_+(K) - Rd(0)]_{ij} = 0$, for $i, j = 1, 2, 3,\ i \neq j$, using the test statistics $Zd(K)$.[16] The results are reported in Table 1.4.[17]

Table 1.4: Lead-lag patterns between weekly size-sorted portfolio returns

		$\widehat{\mathcal{VR}d}_+(K) - \widehat{Rd}(0)$	To		
Lags	S/P	From	small	medium	large
$K = 2$	(1)	small		**0.20** (5.74)*	**0.04** (1.15)
		medium	**0.39** (9.61)*		**0.05** (1.47)
		large	**0.32** (8.21)*	**0.21** (5.42)*	
	(2)	small		**−0.02** (−0.33)	**−0.07** (−1.01)
		medium	**0.20** (3.32)*		**−0.05** (−0.83)
		large	**0.17** (2.74)*	**−0.01** (−0.08)	
$K = 4$	(1)	small		**0.406** (5.42)*	**0.08** (1.14)
		medium	**0.84** (10.39)*		**0.12** (1.756)
		large	**0.67** (9.03)*	**0.41** (5.75)*	
	(2)	small		**−0.00** (−0.00)	**−0.09** (−0.63)
		medium	**0.43** (3.54)*		**−0.05** (−0.38)
		large	**0.34** (2.93)*	**0.04** (0.38)	
$K = 8$	(1)	small		**0.57** (4.11)*	**0.10** (0.73)
		medium	**1.38** (10.21)*		0.18 (1.53)
		large	**1.07** (9.29)*	**0.59** (5.24)*	
	(2)	small		**−0.05** (−0.25)	**−0.16** (−0.72)
		medium	**0.60** (3.28)*		**−0.13** (−0.61)
		large	**0.51** (2.81)*	**0.05** (0.27)	
$K = 16$	(1)	small		**0.54** (2.39)*	**−0.03** (−0.11)
		medium	**1.77** (9.11)*		**0.13** (0.68)
		large	**1.36** (8.42)*	**0.64** (3.80)*	
	(2)	small		**−0.21** (−0.62)	**−0.28** (−0.83)
		medium	**0.67** (2.45)*		**−0.26** (−0.86)
		large	**0.61** (2.22)*	**−0.03** (−0.10)	

[16] For testing $[\mathcal{VR}d_+(K) - Rd(0)]_{ij} = 0$, the $Zd(K)$ statistics are calculated by setting e_f as a column vector that is 1 at the $d(j-1) + i$ entry and 0 otherwise.

[17] In this examination, we divide the whole sample into two sub-samples: 62:07:06-94:12:23 and 94:12:30-13:12:27.

Test statistics in parentheses marked with asterisks indicate that null hypothesis is rejected at 5% level of significance. S/P means the Sample Period; (1): 1962:07:06—1994:12:23 and (2): 1994:12:30—2013:12:27

The results suggest there are strong lead-lag relationships, where medium and large firms lead and small firms lag for all horizons for both sample periods, although the evidence attenuates in the later period, especially at the longer horizon. Nevertheless, there is statistical significance at the 5% level in all such cases. The sign of these terms are all positive and increase with horizon. Also, the size of the coefficients decreases substantially in the later sample period. The evidence is weaker for cross-autocorrelation between current returns of medium sized firms and past returns of small and large ones. We do find that there is evidence of such relationships in the earlier sample period. However, in the later period none of these effects is significant. Finally, with regard to cross-autocorrelation between current returns of large firms and past returns of small and medium sized ones, in no period do we find evidence of this.[18] These results may be interpreted as being consistent with the explanations given in Campbell, Lo and Mackinlay (1997). This is also inconsistent with the random walk hypothesis, but the declining statistical significance may be consistent with improvements in the efficiency/reduction in microstructure effects of these markets.

We also check if the lead-lag patterns are asymmetric. We test the null hypotheses that $[\mathcal{VR}d_+(K) - Rd(0)]_{ij} - [\mathcal{VR}d_+(K) - Rd(0)]_{ji} = 0$, for $i, j = 1, 2, 3,\ i > j$, using the test statistics $Zd(K)$.[19] The results are reported in Table 1.5.

Table 1.5: Asymmetry of lead-lag patterns

		$\left[\widehat{\mathcal{VR}d}_+(K) - \widehat{Rd}(0)\right]_{ij} - \left[\widehat{\mathcal{VR}d}_+(K) - \widehat{Rd}(0)\right]_{ji}$		
Lags	S/P	$(S \to M) - (M \to S)$	$(S \to L) - (L \to S)$	$(M \to L) - (L \to M)$
$K = 2$	(1)	**−0.19** $(-8.75)^*$	**−0.28** $(-8.58)^*$	**−0.16** $(-8.10)^*$
	(2)	**−0.22** $(-6.62)^*$	**−0.23** $(-6.38)^*$	**−0.05** $(-2.31)^*$
$K = 4$	(1)	**−0.44** $(-9.63)^*$	**−0.59** $(-8.68)^*$	**−0.29** $(-7.46)^*$
	(2)	**−0.43** $(-7.15)^*$	**−0.43** $(-6.32)^*$	**−0.09** $(-2.37)^*$
$K = 8$	(1)	**−0.81** $(-10.58)^*$	**−0.97** $(-8.98)^*$	**−0.40** $(-7.02)^*$
	(2)	**−0.68** $(-7.19)^*$	**−0.67** $(-5.79)^*$	**−0.17** $(-3.00)^*$
$K = 16$	(1)	**−1.23** $(-10.16)^*$	**−1.38** $(-8.18)^*$	**−0.51** $(-6.05)^*$
	(2)	**−0.88** $(-6.26)^*$	**−0.89** $(-5.27)^*$	**−0.23** $(-3.03)^*$

S is for small-size portfolio, M is for medium-size portfolio, and L is for large-size portfolio. Test statistics marked with asterisks indicate that the lead-lag relationship is statistically asymmetric at 5%

[18] This test is related to the Granger noncausality test proposed in Pierce and Haugh (1977), where the series are prewhitened before testing zero cross-autocorrelation.

[19] For testing $[\mathcal{VR}d_+(K) - Rd(0)]_{ij} - [\mathcal{VR}d_+(K) - Rd(0)]_{ji} = 0$, the $Zd(K)$ statistics are calculated by setting e_f as a column vector that is 1 at the $d(j-1)+i$ entry, -1 at the $d(i-1)+j$ entry and 0 otherwise.

level of significance. S/P means the Sample Period; (1): 1962:07:06—1994:12:23 and (2): 1994:12:30—2013:12:27

These results can be compared with Campbell, Lo and Mackinlay (1997, page 71, Table 2.9) who look at the asymmetry of the cross-autocorrelation matrices. We find the same direction of asymmetry consistent with their results. The statistical significance does decline in the second period, but is still quite strong.

(3) Multivariate Tests

The above univariate variance ratio tests (Table 1.3A,B,C) provide evidence of linear predictability in returns for small and medium-size portfolios. We next test for the absence of serial correlation in the whole return vector of three size-sorted portfolios, based on univariate parameters derived from the variance ratio matrices $\mathcal{VR}(K)$ and $\mathcal{VR}d(K)$ of X_t. Specifically, we consider the trace and determinant of these matrices, as well as $CS(K)$, $GMV(K)$, and $\pi_+(K)$. Test results based on these statistics are reported in the following table.

Table 1.6: Multivariate variance ratio tests for weekly size-sorted portfolio returns

	Lags			
	$K=2$	$K=4$	$K=8$	$K=16$
First period: 62:07:06-78:09:29				
$\widehat{CS}(K)-\widehat{CS}(1)$	**0.21**	**0.46**	**0.69**	**0.81**
	$(5.04)^*$	$(5.23)^*$	$(4.15)^*$	$(3.09)^*$
$\widehat{GMV}(K)$	**0.39**	**0.42**	**0.43**	**0.41**
	$(4.30)^*$	$(3.53)^*$	$(2.08)^*$	(1.01)
$\widehat{\pi}(K)$	**0.0209**	**0.0180**	**0.0124**	**0.0065**
	$(5.20)^*$	$(7.10)^*$	$(6.59)^*$	$(5.01)^*$
$\mathrm{tr}(\widehat{\mathcal{VR}}(K))$	**3.61**	**4.16**	**5.22**	**5.44**
	$(6.59)^*$	$(7.79)^*$	$(6.89)^*$	$(4.90)^*$
$\det(\widehat{\mathcal{VR}}(K))$	**1.62**	**2.67**	**3.61**	**3.57**
	$(6.72)^*$	$(8.95)^*$	$(8.10)^*$	$(5.15)^*$
$W_F(K)$	**128.51***	**122.06***	**86.39***	**52.06***

Second period: 78:10:06-94:12:23				
$\widehat{CS}(K)-\widehat{CS}(1)$	**0.19**	**0.38**	**0.59**	**0.65**
	$(3.49)^*$	$(3.72)^*$	$(3.68)^*$	$(2.64)^*$
$\widehat{GMV}(K)$	**0.39**	**0.42**	**0.41**	**0.37**
	$(4.24)^*$	$(3.19)^*$	(1.87)	(0.49)
$\widehat{\pi}(K)$	**0.0210**	**0.0197**	**0.0162**	**0.0119**
	$(4.05)^*$	$(5.99)^*$	$(7.17)^*$	$(6.94)^*$
$\mathrm{tr}(\widehat{\mathcal{VR}}(K))$	**3.46**	**4.27**	**5.33**	**6.45**
	$(5.08)^*$	$(7.31)^*$	$(8.06)^*$	$(7.57)^*$
$\det(\widehat{\mathcal{VR}}(K))$	**1.37**	**1.94**	**2.48**	**2.82**
	$(4.03)^*$	$(5.38)^*$	$(5.11)^*$	$(3.99)^*$
$W_F(K)$	**114.27***	**124.62***	**123.80***	**103.19***
Third period: 94:12:30-13:12:27				
$\widehat{CS}(K)-\widehat{CS}(1)$	**0.04**	**0.11**	**0.14**	**0.08**
	(0.63)	(0.91)	(0.71)	(0.29)
$\widehat{GMV}(K)$	**0.34**	**0.35**	**0.33**	**0.27**
	(0.42)	(0.47)	(-0.14)	(-0.77)
$\widehat{\pi}(K)$	**0.0067**	**0.0090**	**0.0065**	**0.0039**
	$(2.19)^*$	$(3.89)^*$	$(3.36)^*$	$(2.53)^*$
$\mathrm{tr}(\widehat{\mathcal{VR}}(K))$	**3.09**	**3.46**	**3.79**	**4.08**
	(0.87)	$(2.30)^*$	$(2.36)^*$	$(2.03)^*$
$\det(\widehat{\mathcal{VR}}(K))$	**1.03**	**1.28**	**1.38**	**1.36**
	(0.31)	(1.39)	(1.12)	(0.69)
$W_F(K)$	**67.28***	**73.23***	**61.90***	**48.20***

The estimates of statistics are reported in the main rows. Test statistics $[Zd_{CS}(K), Z_{GMV}(K), Zd_{\pi}(K), Z_{tr}(K), Z_{\det}(K)$ as defined in (26-27)] in parentheses marked with asterisks indicate statistically significant at 5% level. $W_F\,(K)$ [defined in (28)] is marked with asterisks if it is larger than 12.592, the 5% critical value of $\chi^2(6)$.

There are some differences of opinion between the measures in the most recent period. Specifically, the momentum profit measure is statistically significant at all horizons, and the trace statistic is significant at horizons $K = 4, 8$, and 16, while the other univariate quantities such as the determinant are not significantly different from their null values. In most cases, the univariate statistics are above their predicted values consistent with the earlier results. Although the momentum profit measure is significant in all three periods, the magnitude of the parameter has reduced substantially. The joint test of all the restrictions is strongly significant in all three periods and for all horizons.

We next check whether our results are driven by the choice of subsamples, which we have chosen to match the choices made by CLM for the purpose of replication and comparison. We carry out a rolling window analysis with a (trailing) window of 500 weeks from the beginning of the sample to the end. Below we show the time series of (standard normal) test statistics $Zd_{CS}(K), Z_{GMV}(K)$ and $Zd_{\pi}(K)$ for $K = 4$. This shows that for $\widehat{GMV}(K)$ and $\widehat{CS}(K)$ the sustained decline in statistical significance happened in the decade ending in 2008, although there was an earlier dip in significance in the decade ending in 1999. The profits measure $\widehat{\pi}(K)$ has shown a slower but equally sustained drop in statistical significance. There are some sudden jumps (both up and down) to the level of this statistic in particular, which may be a cause for concern in practice. The $\widehat{GMV}(K)$ statistic seems less affected by such movements.

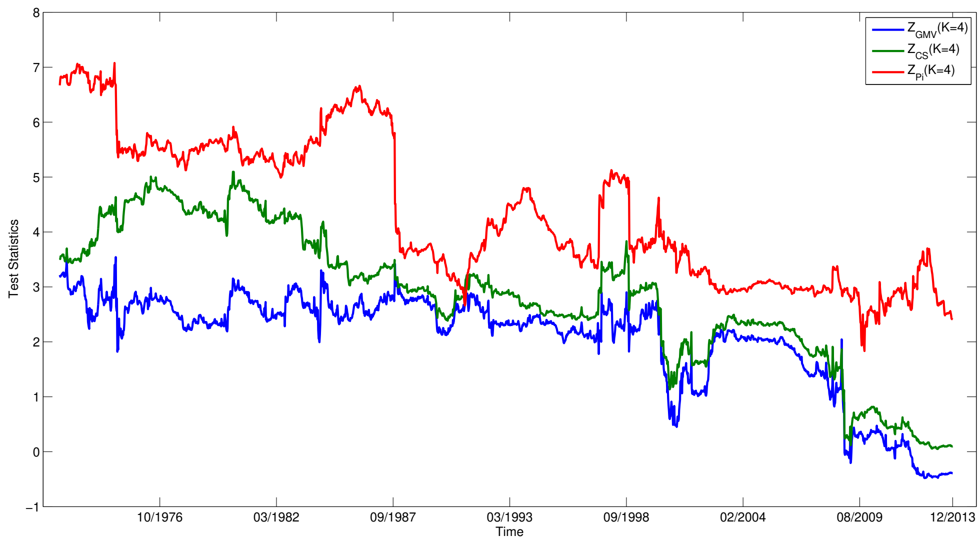


Figure 1.1: Trends of test statistics based on 10 year rolling windows.

1.8.2 Long Horizon

We further investigate the variance ratios at longer horizons. We still use the weekly returns for three size-sorted CRSP portfolios (first, third and fifth quintiles). Here, we work with the bias-corrected estimators

$$\widehat{\mathcal{VR}}^{bc}(K) = \widehat{\mathcal{VR}}(K)\left\{1 + \frac{K-1}{T}\right\} \tag{1.41}$$

$$\widehat{\mathcal{VR}d}^{bc}(K) = \widehat{\mathcal{VR}d}(K)\left\{1 + \frac{K-1}{T}\right\}. \tag{1.42}$$

The ordered eigenvalues may suffer an even larger bias under the null hypothesis, because their limiting distribution is not centred at zero, and so we propose to modify the estimated eigenvalues by

$$\widehat{\lambda}_j^{bc}(K) = \widehat{\lambda}_j\left(\widehat{\mathcal{VR}}^{bc}(K)\right) - \frac{1}{\sqrt{T}} E\lambda_j^*(U_K),$$

where U_K is the limiting matrix distribution of $\sqrt{T}(\widehat{\mathcal{VR}}(K) - I)$. We calculate $E\lambda_j^*(U_K)$ by simulation.

First, we show below in Figure 1.2 the three eigenvalues $\widehat{\lambda}_j^{bc}(K)$ of $\widehat{\mathcal{VR}}^{bc}(K)$ against K for the three sub-samples: the first panel is for $\widehat{\lambda}_j^{bc}(K)$ in the first sub-sample (62:07:06-78:09:29), the second panel shows $\widehat{\lambda}_j^{bc}(K)$ in the second sub-sample (78:10:06-94:12:23) and the third panel shows $\widehat{\lambda}_j^{bc}(K)$ in the third sub-sample (94:12:30-13:12:27). We also use the dashed lines to indicate the 95% pointwise confidence intervals of the largest eigenvalues for each period centred at the null hypothesis. We show out to two years (100 lags), which is quite a long horizon relative to the sample size, and we urge caution in interpreting the results.

We see that the largest eigenvalue increases steadily out to the two year horizon we consider in all three subperiods. In fact, the increase appears to be linear in lag, although the slope is far less than one. The last subperiod has the lowest values throughout, while surprisingly, the second period 1978-1994 seems to have the largest amount of potential linear predictability that could have been exploited during this period. In all cases, the eigenvalues are statistically significant. The apparent increase in predictability at long horizons that this indicates is consistent with the results discussed in the predictive regression literature, see Phillips (2015), in which $X_t(K)$ is regressed on covariates such as (annual) dividend price ratio dated $t-1$.The second and third eigenvalues are quite flat and close to one throughout. This evidence does not seem to be consistent with the fads model, or even the bubble process, although the confidence intervals are quite wide at the longer lags.

We next evaluate the long run behaviour of the $CS(K)$ statistics. Specifically, we consider two one sided statistics:

$$\widehat{CS}_{\pm}(K) = \frac{2}{d(d-1)} \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \left[\widehat{\mathcal{VR}d}_{\pm}^{bc}(K)\right]_{ij}$$

These statistics measure in some average sense the cross dependence for certain directions. We show below the $\widehat{CS}_+(K)$ and $\widehat{CS}_-(K)$ statistics for three weekly size-sorted CRSP portfolio returns against lag K in three sub-samples: the dark solid line is for $\widehat{CS}_+(K)$ in the first sub-sample (62:07:06-78:09:29), the dark dashed line is for $\widehat{CS}_+(K)$ in the second sub-sample (78:10:06-94:12:23), the dark marked line is for $\widehat{CS}_+(K)$ in the third sub-sample (94:12:30-13:12:27); the gray solid line is for $\widehat{CS}_-(K)$ in the first sub-sample,

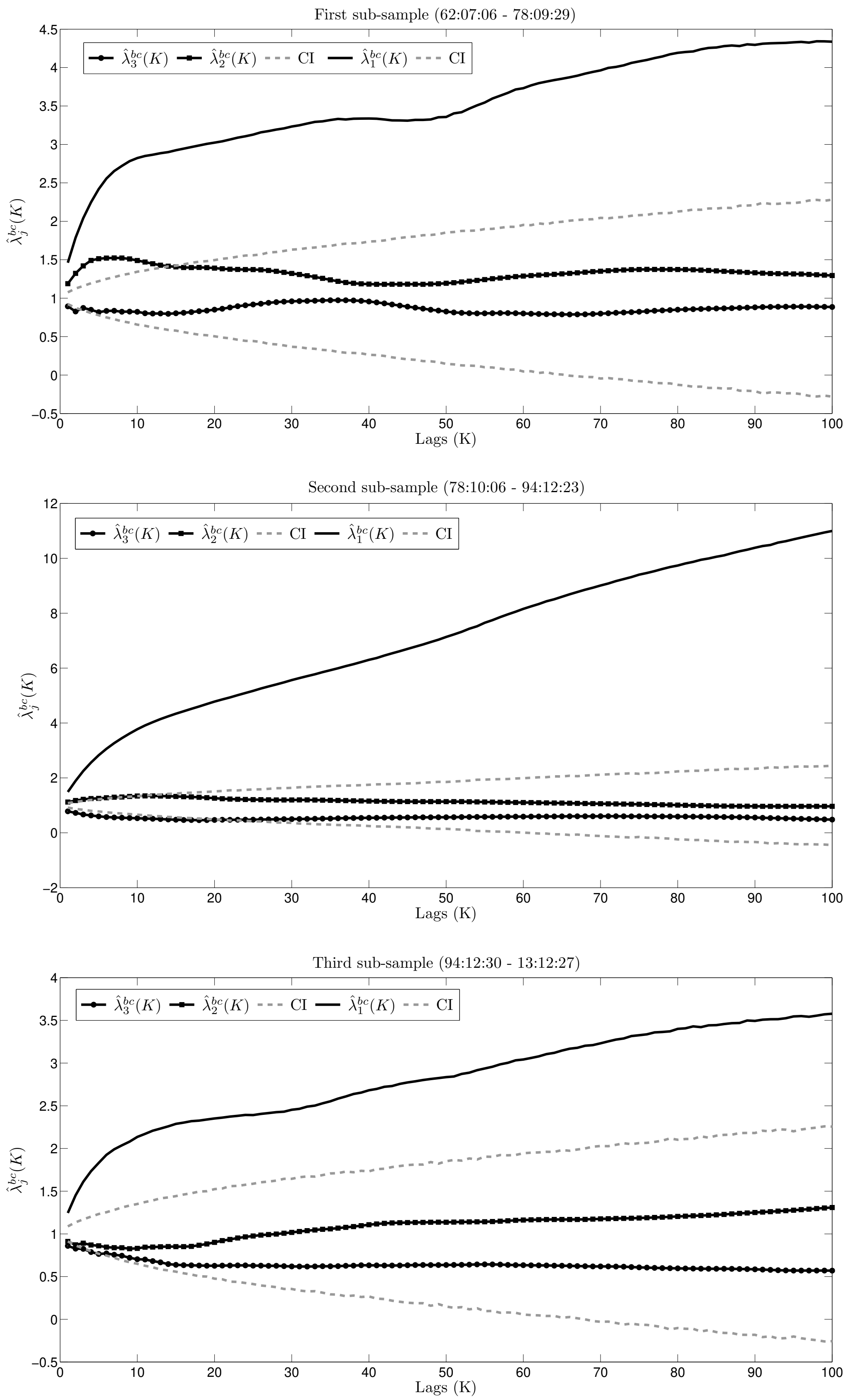


Figure 1.2: The bias corrected eigenvalues of the bias corrected variance ratio matrix in three sub-samples as a function of lags.

the gray dashed line is for $\widehat{CS}_{-}(K)$ in the second sub-sample, and the gray marked line is for $\widehat{CS}_{-}(K)$ in the third sub-sample.

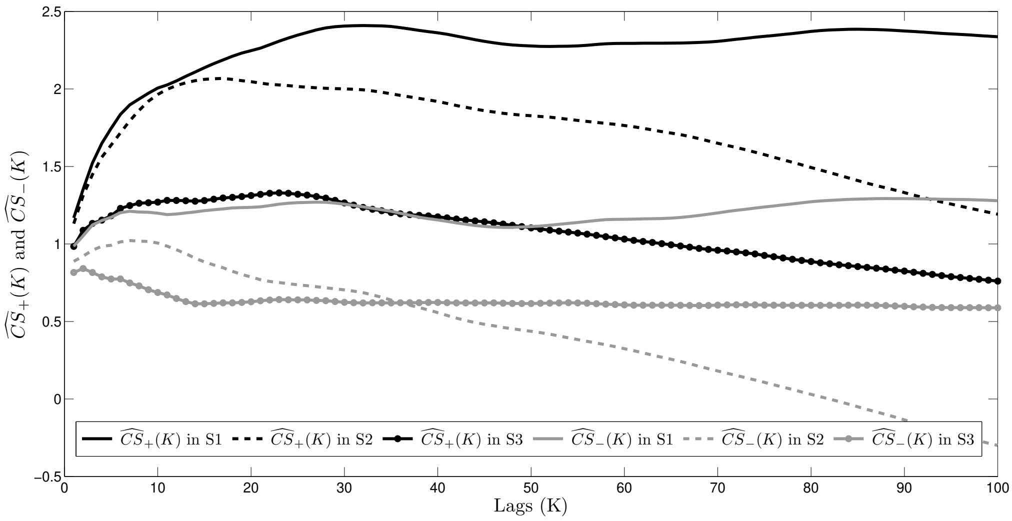


Figure 1.3: $\widehat{CS}_{+}(K)$and $\widehat{CS}_{\text{-}}(K)$ statistics in three sub-samples as a function of lags.

In each subperiod, the $\widehat{CS}_{+}(K)$ measures all exceed the $\widehat{CS}_{-}(K)$ measures over all lags, which means that the average directional cross dependence from larger-size portfolios to smaller-size portfolios are stronger than those in the opposite directions, up to two years. The $\widehat{CS}_{+}(K)$ measures decrease in the recent period over the long horizon. Also the shape of the term structure is quite flat in the most recent period, whereas in the second period, and to a lesser extent in the first period, there seems to be a hump shaped curve suggesting this dependence reaches a maximum somewhere between 10 and 30 weeks. We can further detect that the average statistic, $\widehat{CS}(K) = \left[\widehat{CS}_{+}(K) + \widehat{CS}_{-}(K)\right]/2$, measuring the average cross dependence for both directions between three size-sorted CRSP portfolios, becomes weaker (more efficient) in recent periods at the long horizon.

We then examine the long run $GMV(K)$ statistics. We show below in Figure 1.4 $\widehat{GMV}$ against K in the three sub-samples: the solid line is for $\widehat{GMV}(K)$ in the first sub-sample (62:07:06-78:09:29) and the dashed line is for $\widehat{GMV}(K)$ in the second sub-sample (78:10:06-94:12:23), and the marked line is for $\widehat{GMV}(K)$ in the third sub-sample (94:12:30-13:12:27). For readability we have omitted the confidence intervals, which are quite wide in this case and show that mostly this statistic is consistent with the null hypothesis in the most recent period. In this most recent period there is a quite steep fall off in the statistic out to about 3 months followed by a slower rate of decrease thereafter.

We lastly investigate the $\pi(K)$ statistics. We show below in Figure 1.5 π_{+} against K in three sub-samples: the solid line is for $\widehat{\pi}(K)$ in the first sub-sample (62:07:06-78:09:29) and the dashed line is for $\widehat{\pi}(K)$ in the second sub-sample (78:10:06-94:12:23), and the marked line is for $\widehat{\pi}(K)$ in the third sub-sample (94:12:30-13:12:27). Figure 1.5 shows that

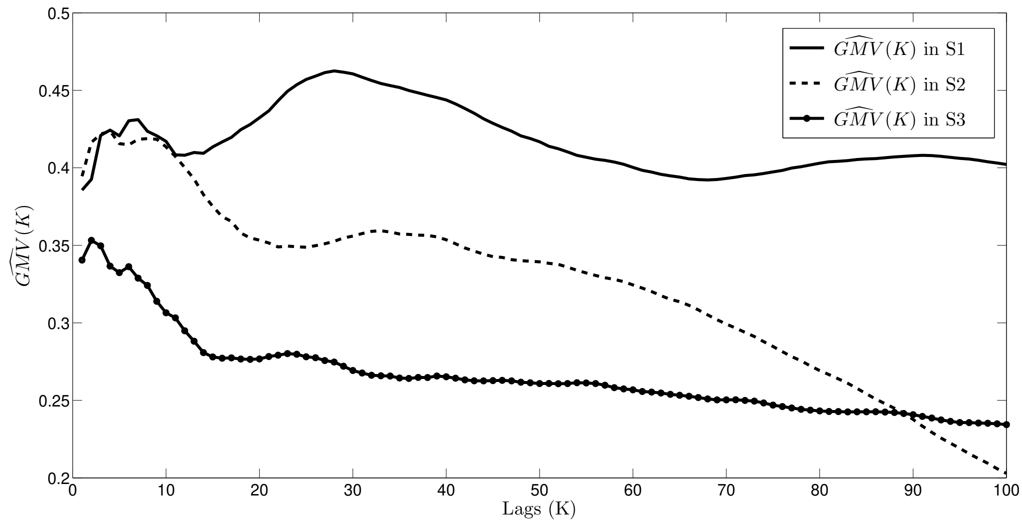


Figure 1.4: $\widehat{GMV}(K)$ statistics in three sub-samples as a function of lags.

the profit measures $\widehat{\pi}(K)$ are positive across all horizons and subsamples we considered (and are also statistically significant for much of the time). We also see that the measures decrease with the lags till around 40 weeks, and then keep at a relatively stable level. In addition, we found the second period 1978-1994 has the largest measures that could have been exploited during this period.

1.9 Conclusions

The first methodological point we make is to propose confidence intervals that are consistent under uncorrelatedness conditions alone and do not require an additional no leverage/symmetric distribution assumption such as maintained in Lo and MacKinlay (1988), CLM, and in much subsequent work. Our confidence intervals are often (although not necessarily so) larger than those used elsewhere, and therefore reduce the significance of any associated test. We believe our theory is more credible with regard to the data generating process we expect for daily or even lower frequency stock returns. The second contribution is about embedding this theory in a multivariate framework. The multivariate variance ratios provides a basis for aggregating the cross correlation behaviour of asset returns and providing tests of the multivariate null hypothesis. It implies many more restrictions on the data than the univariate ratios. We present our theory for a single K and for sequences of K growing. One can also present result for the joint distribution of our test statistics over different horizons, which would provide some control against multiple testing. However, in practice, it is common to consider just a few horizons that have a specific practical meaning, and so there is no real danger of K-snooping here, although this does again add caution to prevent over interpretation.

Our empirical work reports that the US size sorted stock portfolios seem to have come closer to the efficient markets prediction, although there remains some statistically significant linear predictability at the 2 weeks to 16 weeks horizon. Although many of the individual variance ratio statistics do not reject the null hypothesis with our standard errors, the joint tests of the multivariate hypothesis reject at the 1% level in all cases, meaning for all horizons. This is despite the fact that our standard errors are always larger than those of Lo and MacKinlay (1988), which are themselves a lot larger than those based on the iid assumption, which is the world where most applied studies still inhabit.

Typically, three competing explanations are advanced for the predictability in short horizon returns based on past prices (Boudoukh, Richardson, and Whitelaw (1994)): First, microstructure effects such as nonsynchronous trading and bid ask bounce. Second, time varying risk premia reflecting rational behaviour. Third, the irrational behaviour of market participants. It would seem that there is a lot of evidence that microstructure effects have reduced considerably over time. For example, it is hard to find even small cap stocks that do not trade now many times during a day. The microstructure explanation would imply that the long horizon daily or weekly variance ratios should return to unity, but this is not the case in our data even for the most recent period. There is also some evidence that the level (and perhaps therefore the local time variation) of the market risk premium has reduced in recent years, see for example Hertzberg (2010). In the working paper version of this chapter we provided a test of whether the autocorrelations could be explained by time varying risk premia inside a Fama French factor model. We found that this approach could not capture all the linear dependency in the data even in the earlier periods, where the violations were strongest. Therefore, the first two explanations do not seem to be able to match the magnitude of the effects in the earlier periods, although both may make some contribution.[20] On the other hand, the magnitude of the predictability has reduced in the most recent period according to the statistical metrics we have presented here. The long horizon analysis suggests that the largest eigenvalue of the variance ratio matrix grows linearly with horizon, although the slope is far less than the unit slope predicted by the bubble process of Section 1.5.3, which may in principle be consistent with very short bubbly episodes dominated by longer calmer periods.[21]

[20] There is a literature that provides bounds on the implied magnitude of autocorrelations caused by specific microstructure imperfections such as nontrading and a similar literature that provides bounds on the implied magnitude of autocorrelations caused by rational time varying risk premia alone. In both cases strong assumptions are made, see for example Kirby (1998) and Anderson (2011), and Boudoukh, Richardson and Whitelaw (1994), and one "cause" is investigated at a time.

[21] Timmerman (2008) investigates the forecasting performance of a number of linear and nonlinear models and says: "Most of the time the forecasting models perform rather poorly, but there is evidence of relatively short-lived periods with modest return predictability. The short duration of the episodes where return predictability appears to be present and the relatively weak degree of predictability even during such periods makes predicting returns an extraordinarily challenging task". Our (multivariate) evidence does not substantially contradict that; certainly using linear multivariate methods the amount

Alternatively, this may be consistent with a very persistent time varying risk premium of the sort outlined in Section 1.5.4. In any case, the trajectory is flatter (and not statistically significant) in the more recent period, again supporting the claim that market inefficiency has reduced. Although the statistical magnitudes seem to have reduced, it is not clear whether the potential profit from exploiting linear predictability across the whole market has reduced, since the number of tradeable assets has increased and the transactions costs associated with any given trade seem to have reduced, Malkiel (2015).

1.10 Appendix: Proofs of the main results

1.10.1 Proof of Theorem 1.1 and Corollary 1.1

PROOF OF THEOREM 1.1. We first present the proof under Assumption A. For each $j = 1, \ldots, K$,

$$\begin{aligned}\sqrt{T} \cdot \operatorname{vec}\big(\widehat{\Gamma}(j)\big) &= \frac{1}{\sqrt{T}} \sum_{t=j+1}^{T} \left(\widetilde{X}_{t-j} \otimes \widetilde{X}_t\right) - \frac{1}{\sqrt{T}} \sum_{t=j+1}^{T} \widetilde{X}_{t-j} \otimes \big(\overline{X} - \mu\big) \\ &\quad -\big(\overline{X} - \mu\big) \otimes \frac{1}{\sqrt{T}} \sum_{t=j+1}^{T} \widetilde{X}_t + \frac{T-j}{\sqrt{T}} \big(\overline{X} - \mu\big) \otimes \big(\overline{X} - \mu\big) \quad (1.43) \\ &= \frac{1}{\sqrt{T}} \sum_{t=j+1}^{T} \left(\widetilde{X}_{t-j} \otimes \widetilde{X}_t\right) + o_p(1) \quad (1.44)\end{aligned}$$

because $\sum_{t=j+1}^{T} \widetilde{X}_t = O_p(\sqrt{T})$ by the CLT for stationary ergodic martingale difference.

It then follows that

$$\begin{aligned}\sqrt{T}\operatorname{vec}\left(\widehat{\mathcal{VR}}_{+}(K) - I_d\right) &= \sqrt{T} \cdot \sum_{j=1}^{K-1} 2\left(1 - \frac{j}{K}\right) \cdot \operatorname{vec}\left(\widehat{R}(j)\right) \\ &= \big(\Sigma^{-1/2} \otimes \Sigma^{-1/2}\big) \cdot \sum_{j=1}^{K-1} c_j \frac{1}{\sqrt{T}} \sum_{t=j+1}^{T} \widetilde{X}_{t-j} \otimes \widetilde{X}_t + o_p(1) \\ &= \big(\Sigma^{-1/2} \otimes \Sigma^{-1/2}\big) \cdot \frac{1}{\sqrt{T}} \sum_{t=K}^{T} \left[\sum_{j=1}^{K-1} c_j \left(\widetilde{X}_{t-j} \otimes \widetilde{X}_t\right)\right] + o_p(1) \\ &=: \big(\Sigma^{-1/2} \otimes \Sigma^{-1/2}\big) \cdot \frac{1}{\sqrt{T}} \sum_{t=K}^{T} Z_t + o_p(1). \quad (1.45)\end{aligned}$$

This is because, by stationary ergodicity of $\widetilde{X}_t \widetilde{X}_t^\intercal$, the Ergodic theorem and continuous

of predictability we have found and its durability is limited and has reduced over time even through the recent financial crisis.

mapping on $T^{-1}\sum_{t=1}^{T}\widetilde{X}_t\widetilde{X}_t^{\intercal}$ yields $\widehat{\Sigma}^{-1/2}-\Sigma^{-1/2}=o_p(1)$, and consequently for each j

$$\begin{aligned}\operatorname{vec}(\widehat{R}(j)) &= \operatorname{vec}\left(\left[\widehat{\Sigma}^{-1/2}-\Sigma^{-1/2}+\Sigma^{-1/2}\right]\widehat{\Gamma}(j)\left[\widehat{\Sigma}^{-1/2}-\Sigma^{-1/2}+\Sigma^{-1/2}\right]\right)\\ &= \left(\Sigma^{-1/2}\otimes\Sigma^{-1/2}\right)\operatorname{vec}(\widehat{\Gamma}(j))+o_p(1)\end{aligned} \tag{1.46}$$

It now suffices to derive the limiting distribution of Z_t. Take any d^2-dimensional real constant vector $a=(a_1,\ldots,a_{d^2})^{\intercal}$, and note that $a^{\intercal}Z_t$ is a martingale difference sequence. Then, since by Assumption A2

$$E(a^{\intercal}Z_t)^2=a^{\intercal}\operatorname{var}(Z_t)a=a^{\intercal}\left[\sum_{j=1}^{K-1}\sum_{k=1}^{K-1}c_jc_k\Xi_{jk}\right]a<\infty,$$

where $\Xi_{jk}=E[\widetilde{X}_{t-j}\otimes\widetilde{X}_t][\widetilde{X}_{t-k}\otimes\widetilde{X}_t]^{\intercal}$, the CLT for stationary ergodic martingale difference gives

$$a^{\intercal}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}Z_t\right)\Longrightarrow N\left(0,a^{\intercal}\left[\sum_{j=1}^{K-1}\sum_{k=1}^{K-1}c_jc_k\Xi_{jk}\right]a\right). \tag{1.47}$$

completing the proof in view of the Cramér-Wold device, continuous mapping and Slutsky's theorem.

Similar arguments apply when we work with Assumption MH*. We note that the expansion (1.44) for $\sqrt{T}\cdot\operatorname{vec}(\widehat{\Gamma}(j))$ is still valid because the summations in the second, third and fourth terms in (1.43) still converge in probability to zero due to the CLT for mixing sequence, Herrndorf (1985, Theorem 0) whose regularity conditions are satisfied by MH1-MH3. Finally, condition MH2 and MH3 allow for the LLN for mixing variables, White (1984, Corollary 3.48), yielding (1.46) and (1.45) as before.

Now we are only left with verifying (1.47). For any d^2-dimensional constant vector a, $a^{\intercal}Z_t$ preserves the mixing property of $\widetilde{X}_t$ with the same rate, so by Herrndorf's CLT we have

$$a^{\intercal}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}Z_t\right)\Longrightarrow N\left(0,a^{\intercal}\left[\sum_{j=1}^{K-1}\sum_{k=1}^{K-1}c_jc_k\Xi_{jk}\right]a\right),$$

where $\Xi_{jk}=\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}E[\widetilde{X}_{t-j}\otimes\widetilde{X}_t][\widetilde{X}_{t-k}\otimes\widetilde{X}_t]^{\intercal}$. The CLT above holds provided the following regularity conditions are ensured: $E(a^{\intercal}Z_{tj})=0$, $\sup_t E|a^{\intercal}Z_t|^{\beta}<\infty$ for some $\beta>2$, and finally

$$\lim_{T\to\infty}\frac{1}{T}E\left(\sum_{t=1}^{T}a^{\intercal}Z_t\right)^2=\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\operatorname{var}\left(a^{\intercal}Z_t\right)=a^{\intercal}\left[\sum_{j=1}^{K-1}\sum_{k=1}^{K-1}c_jc_k\Xi_{jk}\right]a$$

is positive and finite. The first condition is trivial by MH1, and the second and third conditions are satisfied by MH2, MH3 and positive definiteness of $Q(K)$. The desired result readily follows. The arguments for the diagonally normalized is identical everywhere except that we have $\text{vec}(\widehat{R}d(j)) = \big(D^{-1/2} \otimes D^{-1/2}\big)\,\text{vec}(\widehat{\Gamma}(j)) + o_p(T^{-1/2})$ instead of (1.46). The entire proof is now complete. ■

Proof of Corollary 1.1. It suffices to show consistency of $\widehat{\Xi}_{jk}$ for each j and k. Writing

$$\widehat{\Xi}_{jk} = \frac{1}{T}\sum_{t=\max\{j,k\}+1}^{T}\left[\left(\widetilde{X}_{t-j}\otimes\widetilde{X}_t\right)\left(\widetilde{X}_{t-k}\otimes\widetilde{X}_t\right)^{\intercal}\right] + o_p(1).$$

we see that the desired result follows by applying either the Ergodic theorem or the Law of Large Numbers for mixing variables depending upon the set of assumption being imposed. The regularity conditions for each theorem are ensured by Assumption A2 and MH3, respectively. ■

1.10.2 Proof of Theorem 1.2

Proof. For later reference, we first derive the limiting distribution of

$$\frac{1}{\sqrt{T}}\sum_{t=K}^{T}\left[\frac{1}{\sqrt{K}}\sum_{j=1}^{K-1}c_j(\widetilde{X}_{t-j}\otimes\widetilde{X}_t)\right] =: \frac{1}{\sqrt{T}}\sum_{t=K}^{T}Z_{Tt}. \tag{1.48}$$

The asymptotic normality is established by applying the central limit theorem for triangular arrays of martingale difference in Pollard (1984, page 171) on $\{Z_{Tt},\mathcal{F}_{Tt}\}$, where $\mathcal{F}_{Tt} = \mathcal{F}_t = \sigma(\widetilde{X}_s;\ s \le t)$. Specifically, for some arbitrary non-zero constant vector $a = (a_1, a_2, \ldots, a_{d^2})^{\intercal}$ we check the following conditions:

(i) $T^{-1}\sum_t E((a^{\intercal}Z_{Tt})^2|\mathcal{F}_{t-1}) \xrightarrow{P} a^{\intercal}\eta a > 0$

(ii) $\forall\varepsilon > 0$, $T^{-1}\sum_t E((a^{\intercal}Z_{Tt})^2 1\{|a^{\intercal}Z_{Tt}| > \varepsilon\sqrt{T}\}|\mathcal{F}_{t-1}) \xrightarrow{P} 0$

under which (via the Cramér-Wold Theorem) it will follow that

$$\frac{1}{\sqrt{T}}\sum_{t=K}^{T}Z_{Tt} \Longrightarrow N\big(0,\eta\big). \tag{1.49}$$

As for the first condition (i), it suffices to show

$$\frac{1}{T}\sum_{t=K}^{T}E(a^{\intercal}Z_{Tt})^2 \to a^{\intercal}\eta a \tag{1.50}$$

$$\frac{1}{T}\sum_{t=K}^{T}E\big((a^{\intercal}Z_{Tt})^2|\mathcal{F}_{t-1}\big) - E(a^{\intercal}Z_{Tt})^2 \xrightarrow{P} 0. \tag{1.51}$$

We first consider (1.50); denote by $\widetilde{X}_{it}$ the ith element of the d-vector $\widetilde{X}_t$.

Then the moment-cumulant relationship formula (see e.g. Hannan (1970, page 23)) suggests that for any $q, w, e, u = 1, \ldots, d$

$$
\begin{aligned}
&\frac{1}{K}\sum_{j=1}^{K-1}\sum_{r=1}^{K-1} c_j c_r E\big(\widetilde{X}_{q,t-j}\widetilde{X}_{w,t-r}\widetilde{X}_{et}\widetilde{X}_{ut}\big) \\
&= \frac{1}{K}\sum_{j=1}^{K-1}\sum_{r=1}^{K-1} c_j c_r \Big[\gamma_{qw}(j-r)\gamma_{eu}(0) + \gamma_{qe}(j)\gamma_{wu}(r) + \gamma_{qu}(j)\gamma_{we}(r) + \kappa_{qweu}(-j,-r,0,0)\Big] \\
&= \gamma_{qw}(0)\gamma_{eu}(0)\frac{1}{K}\sum_{p=1}^{K-1} 4\Big(1-\frac{p}{K}\Big)^2 + \frac{1}{K}\sum_{j=1}^{K-1}\sum_{r=1}^{K-1}\kappa_{qweu}(-j,-r,0,0) \;\to\; \frac{4}{3}\cdot\gamma_{qw}(0)\gamma_{eu}(0)
\end{aligned}
$$

where $\gamma_{qw}(u) = E(\widetilde{X}_{qt}\widetilde{X}_{w,t-u})$. This is due to uncorrelatedness of $\widetilde{X}_t$ and double summability of the fourth order cumulant κ (Assumption S). Consequently,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=K}^{T} E\big[(a^\intercal Z_{Tt})\big]^2 &= \frac{1}{T}\sum_{t=K}^{T} E\left[\frac{1}{\sqrt{K}}\sum_{j=1}^{K-1} a^\intercal c_j(\widetilde{X}_{t-j}\otimes\widetilde{X}_t)\right]^2 \\
&= \left(\frac{T-K}{T}\right) a^\intercal\left[\frac{1}{K}\sum_{j=1}^{K-1}\sum_{r=1}^{K-1} c_j c_r E\big(\widetilde{X}_{t-j}\widetilde{X}_{t-r}^\intercal\otimes\widetilde{X}_t\widetilde{X}_t^\intercal\big)\right] a \\
&= \frac{4}{3}a^\intercal(\Sigma\otimes\Sigma)a + O\left(\frac{K}{T}\right) \to a^\intercal\eta a
\end{aligned}
$$

which is strictly positive by the assumption that Σ is positive definite. This finite limit also implies (1.51) as a consequence of the ergodic theorem in view of the inherited stationary ergodicity through a measurable mapping, see for example Karlin and Taylor (1975, page 487-488).

It remains to check the conditional Lindeberg condition (ii). We have for any $\epsilon > 0$,

$$
\begin{aligned}
&P\left(\left|\frac{1}{T}\sum_{t=1}^{T} E\left((a^\intercal Z_{Tt})^2 1_{\{|a^\intercal Z_{Tt}|>\varepsilon\sqrt{T}\}}|\mathcal{F}_{t-1}\right)\right| > \epsilon\right) \\
&\le \frac{1}{\epsilon}E\left|\frac{1}{T}\sum_{t=1}^{T} E\left((a^\intercal Z_{Tt})^2 1_{\{|a^\intercal Z_{Tt}|>\varepsilon\sqrt{T}\}}|\mathcal{F}_{t-1}\right)\right| \le \frac{1}{\epsilon}\left(\frac{1}{\varepsilon\sqrt{T}}\right)^{\frac{\delta}{2}} E\left((a^\intercal Z_{Tt})^{2+\frac{\delta}{2}}\right) \\
&\le \frac{1}{\epsilon}\left(\frac{1}{\varepsilon\sqrt{T}}\right)^{\frac{\delta}{2}} E\left(\sum_{i=1}^{d^2}\frac{1}{\sqrt{K}}\sum_{j=1}^{K-1} c_j a_i \widetilde{X}_{\lceil i/d\rceil,0}\widetilde{X}_{i-d(\lceil i/d\rceil-1),j}\right)^{2+\frac{\delta}{2}} \\
&\le \frac{1}{\epsilon}\left(\frac{1}{\varepsilon\sqrt{T}}\right)^{\frac{\delta}{2}}\left(\frac{1}{\sqrt{K}}\sum_{i=1}^{d^2}|a_i|\left[E\left|\sum_{j=1}^{K-1}\widetilde{X}_{i-d(\lceil i/d\rceil-1),j}\widetilde{X}_{\lceil i/d\rceil,0}\right|^{2+\delta/2}\right]^{1/(2+\delta/2)}\right)^{2+\frac{\delta}{2}} \\
&= \frac{C}{T^{\delta/4}}\left(\frac{1}{\sqrt{K}}\right)^{2+\delta/2}\max\{K^{1+\delta/4}, K\} = O\left(\frac{1}{T^{\delta/4}}\right)
\end{aligned}
$$

where $\lceil \cdot \rceil$ is the ceiling function. Here we used stationarity, law of total expectations, the moment condition $E|\widetilde{X}_t|^{4+\delta} \leq C$ (where δ is as in Assumption A$'$), Minkowski's inequality, and Burkholder's inequality for martingale difference (e.g. Gut (2005, page 506-507)).

Now, starting from the decomposition (42) in the proof of Theorem 1.1, we can easily see using Chebyshev's inequality, results from Theorem 1.1 and elementary algebra that

$$\begin{aligned}
\sqrt{\frac{T}{K}}\mathrm{vec}\Big(\widehat{\mathcal{VR}}(K)_+ - I_d\Big) &= \frac{\sqrt{T}}{\sqrt{K}} \cdot \sum_{j=1}^{K-1} 2\left(1-\frac{j}{K}\right) \cdot \mathrm{vec}\left(\widehat{R}(j)\right) \\
&= (\widehat{\Sigma}^{-1/2} \otimes \widehat{\Sigma}^{-1/2}) \frac{1}{\sqrt{K}} \sum_{j=1}^{K-1} c_j \Bigg\{ \left[\frac{1}{\sqrt{T}} \sum_{t=j+1}^{T} \widetilde{X}_{t-j} \otimes \widetilde{X}_t \right] \\
&\quad - \left[\frac{1}{\sqrt{T}} \sum_{t=j+1}^{T} \widetilde{X}_{t-j} \otimes \left(\overline{X} - \mu\right) \right] - \left[\left(\overline{X} - \mu\right) \otimes \frac{1}{\sqrt{T}} \sum_{t=j+1}^{T} \widetilde{X}_t \right] \\
&\quad + \left[\frac{T-j}{\sqrt{T}} \left(\overline{X} - \mu\right) \otimes \left(\overline{X} - \mu\right) \right] \Bigg\} \\
&= (\widehat{\Sigma}^{-1/2} \otimes \widehat{\Sigma}^{-1/2}) \frac{1}{\sqrt{T}} \sum_{t=K}^{T} \left[\frac{1}{\sqrt{K}} \sum_{j=1}^{K-1} c_j (\widetilde{X}_{t-j} \otimes \widetilde{X}_t) \right] + O_P\left(\sqrt{\frac{K}{T}}\right) + o_p\left(\sqrt{\frac{K}{T}}\right) \\
&= (\Sigma^{-1/2} \otimes \Sigma^{-1/2}) \frac{1}{\sqrt{T}} \sum_{t=K}^{T} Z_{Tt} + o_P(1) \qquad (1.52)
\end{aligned}$$

Now the desired asymptotic distribution holds in view of the results above and consistency of standard error via the ergodic theorem, completing the proof. ■

1.10.3 Proof of Theorem 1.3

PROOF. The proof proceeds by showing asymptotic equivalence of the trace (of the multivariate variance ratio) test and the likelihood ratio (LR) test under the null and alternative hypotheses. That is,

$$f\left(\mathrm{tr}\left(\widehat{\mathcal{VR}}(K)\right)\right) - LR \xrightarrow{P} 0 \qquad (1.53)$$

for some function f, in which case the tests based on two statistics will possess the same large sample properties.

Recall the alternative estimator $\widehat{\mathcal{VR}}^{\&}(K)$. From the definitions it can be readily shown that

$$\widehat{\mathcal{VR}}(K) - \widehat{\mathcal{VR}}^{\&}(K) = \frac{1}{K}\sum_{r=1}^{K-2}\left\{\widehat{\Sigma}^{-1/2}\left[(K-r)\frac{1}{T}\sum_{t=r+1}^{K-1}\left(X_t - \overline{X}\right)\left(X_{t-r} - \overline{X}\right)^{\intercal}\right]\widehat{\Sigma}^{-1/2}\right\}$$
$$+\frac{1}{K}\sum_{r=1}^{K-2}\left\{\widehat{\Sigma}^{-1/2}\left[(K-r)\frac{1}{T}\sum_{t=2}^{K-1}\left(X_{t-r} - \overline{X}\right)\left(X_t - \overline{X}\right)^{\intercal}\right]\widehat{\Sigma}^{-1/2}\right\} + o_p(1) \quad (1.54)$$

converges in probability to zero because each term in square brackets is $o_p(1)$ by Chebyshev's inequality and $\widehat{\Sigma}^{-1/2} - \Sigma^{-1/2} = o_p(1)$.

Now that we have $f(\mathrm{tr}(\widehat{\mathcal{VR}}(K))) - f(\mathrm{tr}(\widehat{\mathcal{VR}}^{\&}(K))) = o_p(1)$ due to linearity of trace, it remains to show that

$$f\left(\mathrm{tr}\left(\widehat{\mathcal{VR}}^{\&}(K)\right)\right) - LR \xrightarrow{P} 0.$$

Let the 'coefficient matrix' Φ be the matrix of ones except for the $(T-K)\times(T-K)$ triangular blocks in the northeast and southwest corners where the entries are all zero. Then denoting by i a conformable column vector of ones, we have

$$\widehat{\Sigma}(K) = \frac{1}{T}\left(\Phi X - \Phi i\overline{X}^{\intercal}\right)^{\intercal}\left(\Phi X - \Phi i\overline{X}^{\intercal}\right) = \frac{1}{T}\left(X - i\overline{X}^{\intercal}\right)^{\intercal}\Phi^{\intercal}\Phi\left(X - i\overline{X}^{\intercal}\right)$$

from which it follows that

$$\widehat{\mathcal{VR}}^{\&}(K) = \frac{1}{K}[(A^{\intercal}A)^{-1/2}]\cdot[A^{\intercal}\Phi^{\intercal}\Phi A]\cdot[(A^{\intercal}A)^{-1/2}]$$

where $A := \left(X - i\overline{X}^{\intercal}\right)$.

The rejection region based on the likelihood ratio statistic is given by

$$LR = \log\left(\frac{\det\Big[\left(X - i\widehat{\mu}_1^{\intercal}\right)^{\intercal}\Sigma_{q*}^{-1}\left(X - i\widehat{\mu}_1^{\intercal}\right)\Big]}{\det\Big[\left(X - i\overline{X}^{\intercal}\right)^{\intercal}\left(X - i\overline{X}^{\intercal}\right)\Big]}\right) < k$$

for some positive threshold constant k, where $\widehat{\mu}_1 \equiv \widetilde{X}$ is the maximum likelihood estimate of the mean $\mu = EX_t$ under the *alternative* hypotheses. Using a standard property of the logarithmic determinant we see that

$$\begin{aligned} LR &= \log\left(\det\left\{\Big[\left(X - i\overline{X}^{\intercal}\right)^{\intercal}\left(X - i\overline{X}^{\intercal}\right)\Big]^{-1}\Big[\left(X - i\widetilde{X}^{\intercal}\right)^{\intercal}\Sigma_{q*}^{-1}\left(X - i\widetilde{X}^{\intercal}\right)\Big]\right\}\right) \\ &\leq \mathrm{tr}\left(\Big[\left(X - i\overline{X}^{\intercal}\right)^{\intercal}\left(X - i\overline{X}^{\intercal}\right)\Big]^{-1}\Big[\left(X - i\widetilde{X}^{\intercal}\right)^{\intercal}\Sigma_{q*}^{-1}\left(X - i\widetilde{X}^{\intercal}\right)\Big] - I\right) \\ &\leq \mathrm{tr}\left(\widehat{\Sigma}^{-1}\cdot\frac{1}{T}\Big[\left(X - i\widetilde{X}^{\intercal}\right)^{\intercal}\Sigma_{q*}^{-1}\left(X - i\widetilde{X}^{\intercal}\right)\Big]\right). \end{aligned} \quad (1.55)$$

Besides, it follows by the cyclic property of the trace operator that

$$
\begin{aligned}
\operatorname{tr}\left(\widehat{\mathcal{VR}}^{\&}(K)\right) &= \frac{1}{K}\operatorname{tr}\Big([(A^{\intercal}A)^{-1}]\cdot[A^{\intercal}\Phi^{\intercal}\Phi A]\Big) \\
&= \frac{1}{K}\operatorname{tr}\left(T\left[\left(X-i\overline{X}^{\intercal}\right)^{\intercal}\left(X-i\overline{X}^{\intercal}\right)\right]^{-1}\cdot\frac{1}{T}\left[\left(X-i\overline{X}^{\intercal}\right)^{\intercal}\Phi^{\intercal}\Phi\left(X-i\overline{X}^{\intercal}\right)\right]\right) \\
&= \frac{1}{K}\operatorname{tr}\left(\widehat{\Sigma}^{-1}\cdot\frac{1}{T}\left[\left(X-i\widetilde{X}^{\intercal}+i\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right)^{\intercal}\Phi^{\intercal}\Phi\left(X-i\widetilde{X}^{\intercal}+i\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right)\right]\right).
\end{aligned}
$$

Now multiplying the last quantity by the horizon K, $q>0$, adding $d = tr(I_d)$, and then lastly multiplying by some constant $\alpha > 0$ give

$$
\begin{aligned}
&\operatorname{tr}\left(\widehat{\Sigma}^{-1}\cdot\frac{1}{T}\left[\left(X-i\widetilde{X}^{\intercal}+i\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right)^{\intercal}\{\alpha(I+q\Phi^{\intercal}\Phi)\}\left(X-i\widetilde{X}^{\intercal}+i\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right)\right]\right) \\
&= \operatorname{tr}\left(\widehat{\Sigma}^{-1}\cdot\frac{1}{T}\left[\left(X-i\widetilde{X}^{\intercal}+i\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right)^{\intercal}\{\Sigma_q^{-1}+0^*\}\left(X-i\widetilde{X}^{\intercal}+i\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right)\right]\right)
\end{aligned} \tag{1.56}
$$

where 0^* is the matrix of zeros except for the $(K-1)\times(K-1)$ blocks in the northwest and southeast corners. The reader is directed to Faust (1992, Lemma 1) for the proof of the equivalence relationship $\alpha(I+q\Phi^{\intercal}\Phi)\equiv\Sigma_q^{-1}+0^*$. Now replacing the sample estimator for the cross-sectional variance by its population version (with some negligible error), we see that the difference between (1.56) and (1.55) multiplied by $\sqrt{T}$ is given by

$$
\begin{aligned}
&\sqrt{T}\cdot\operatorname{tr}\left(\Sigma^{-1}\cdot\frac{1}{T}\left[\left(i\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right)^{\intercal}\cdot\Sigma_q^{-1}\cdot\left(i\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right)\right]\right)+o_p(1) \\
&\quad= \operatorname{tr}\left(\Sigma^{-1}\cdot\left[\sqrt{T}\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)^{\intercal}\left\{\frac{i^{\intercal}\cdot\Sigma_q^{-1}\cdot i}{T}\right\}\left(\widetilde{X}^{\intercal}-\overline{X}^{\intercal}\right)\right]\right)+o_p(1)
\end{aligned}
$$

because the trace is a linear mapping. It is trivial to show that the term inside $\{\cdot\}$ is bounded in probability. Further, the proof of Proposition 2 in Faust (1992) suggests that the individual entries of the squared bracket converges in probability to zero (hence so does the entire matrix), yielding

$$
\sqrt{T}\left|\alpha\left\{d+qK\cdot\operatorname{tr}\left(\widehat{\mathcal{VR}}^{\&}(K)\right)\right\}-LR\right|\xrightarrow{p}0. \tag{1.57}
$$

This suggests that there exist some α and q for which the trace test has the same large sample properties of the LR test against the $\phi-$best class alternatives. Since the sequence of the LR tests with $q^*=\delta/\sqrt{T}$ is locally most powerful invariant, e.g. Engle (1984), the proof is complete. ■

1.10.4 Proof of Theorems 1.4, 1.5, 1.6 and 1.7, and (1.36)

Proof of Theorem 1.4. Consider the K period returns $X_t(K) = K\mu + p_t - p_{t-K} = \sum_{s=t-K}^{t} \varepsilon_s + \sum_{s=t-K}^{t}(\eta_s - \eta_{s-1}) = K\mu + \sum_{s=t-K}^{t} \varepsilon_s + \eta_t - \eta_{t-K}$. These have variance

$$\begin{aligned}
\Sigma_K &= \mathrm{var}(X_t(K)) = \mathrm{var}\left(\sum_{s=t-K}^{t} \varepsilon_s\right) + \mathrm{var}\left(\eta_t - \eta_{t-K}\right) \\
&= KE\varepsilon_s\varepsilon_s^{\intercal} + E\left((\eta_t - \eta_{t-K})(\eta_t - \eta_{t-K})^{\intercal}\right) = K\Omega_\varepsilon + \Omega_\eta(K),
\end{aligned}$$

where $\Omega_\eta(k) = \mathrm{var}\left(\eta_t - \eta_{t-k}\right) \geq 0$, $k = 1, 2, \ldots$. Therefore, $\mathcal{VR}(K) = \Sigma_1^{-1/2}\Sigma_K\Sigma_1^{-1/2}/K$ and $\mathcal{VR}d(K) = D_1^{-1/2}\Sigma_K D_1^{-1/2}/K$. Note that as $K \to \infty$, $\Omega_\eta(K) \to 2\Omega_\eta = 2\mathrm{var}\left(\eta_t\right)$. It follows that as $K \to \infty$

$$\begin{aligned}
\mathcal{VR}(K) &= K^{-1}\Sigma_1^{-1/2}\Sigma_K\Sigma_1^{-1/2} = K^{-1}\Sigma_1^{-1/2}\left(K\Omega_\varepsilon + \Omega_\eta(K)\right)\Sigma_1^{-1/2} \\
&\to \Sigma_1^{-1/2}\Omega_\varepsilon\Sigma_1^{-1/2} = \Sigma_1^{-1/2}\left[\Sigma_1 - \Omega_\eta(1)\right]\Sigma_1^{-1/2} \\
&= I - \Sigma_1^{-1/2}\Omega_\eta(1)\Sigma_1^{-1/2} \leq I,
\end{aligned}$$

since Σ_1 and $\Omega_\eta(1)$ are positive semidefinite. The strict inequality holds since $\Omega_\eta(1)$ is assumed strictly positive definite. By similar arguments,

$$\begin{aligned}
\mathcal{VR}d(K) &= K^{-1}D_1^{-1/2}\Sigma_K D_1^{-1/2} = K^{-1}D_1^{-1/2}\left(K\Omega_\varepsilon + \Omega_\eta(k)\right)D_1^{-1/2} \\
&\to D_1^{-1/2}\Omega_\varepsilon D_1^{-1/2} = D_1^{-1/2}\left(\Sigma_1 - \Omega_\eta(1)\right)D_1^{-1/2} \\
&= D_1^{-1/2}\Sigma_1 D_1^{-1/2} - D_1^{-1/2}\Omega_\eta(1)D_1^{-1/2} \\
&= Rd\left(0\right) - D_1^{-1/2}\Omega_\eta(1)D_1^{-1/2} \leq Rd\left(0\right)
\end{aligned}$$

which is the instantaneous correlation matrix of the return process. ■

Proof of Theorem 1.5. This follows from the multivariate extension of Theorem 1 of Liu and Wu (2010) applied to the frequency $\theta = 0$. The weighting scheme automatically satisfies their condition 1. See also Andrews (1991). ■

Proof of (1.36). For simplicity we suppose that $p_t = \delta_T p_{t-1} + \varepsilon_t$ with ε_t iid with variance σ_ε^2 and $\delta_T = 1 + \frac{c}{k_T}$, where $k_T = T^\alpha, \alpha \in (0, 1/2)$ and some positive constant c. According to Phillips and Magdalinos (2007, Theorem 4.3) we have

$$\left((\delta_T^{-T}/k_T)\sum_{t=1}^{T} p_{t-1}\varepsilon_t, (\delta_T^{-2T}/k_T^2)\sum_{t=1}^{T} p_{t-1}^2\right) \Longrightarrow (XY, Y^2),$$

where X, Y are iid copies of a $N(0, \sigma_\varepsilon^2/2c)$ distribution.

Since the observed return X_t is the difference of the log prices we have $X_t = p_t - p_{t-1} =$

$\frac{c}{k_T}p_{t-1} + \varepsilon_t$, and consequently the sum of the squared return is

$$\begin{aligned}
\sum_{t=1}^{T} X_t^2 &= \frac{c^2}{k_T^2}\sum_{t=1}^{T} p_{t-1}^2 + \frac{2c}{k_T}\sum_{t=1}^{T} p_{t-1}\varepsilon_t + \sum_{t=1}^{T}\varepsilon_t^2 \\
&\Rightarrow \frac{c^2}{k_T^2}k_T^2\delta_T^{2T}Y^2 + \frac{2c}{k_T}k_T\delta_T^T XY + T\sigma_\varepsilon^2 \\
&= c^2\delta_T^{2T}Y^2 + R,
\end{aligned}$$

where R is a generic remainder term that contains smaller order terms. The first term dominates the others because $\delta_T^{2T} = (1 + \frac{c}{k_T})^{2T} \to \infty$ very fast. Therefore, we have

$$\delta_T^{-2T}\sum_{t=1}^{T} X_t^2 \Longrightarrow c^2Y^2. \tag{1.58}$$

Likewise,

$$X_t(2) = p_t - p_{t-2} = (\delta_T^2 - 1)p_{t-2} + \varepsilon_t + \delta_T\varepsilon_{t-1} \simeq \frac{2c}{k_T}p_{t-2} + \varepsilon_t + \delta_T\varepsilon_{t-1},$$

by the Binomial approximation because $c/k_T = c/T^\alpha$ becomes negligible as T gets bigger.

Therefore,

$$\delta_T^{-2T}\sum_{t=1}^{T} X_t(2)^2 \Longrightarrow 4c^2Y^2.$$

Similarly for general K, as $T \to \infty$ we have:

$$X_t(K) = \left(\delta_T^K - 1\right)p_{t-K} + \sum_{j=0}^{K-1}\delta_T^j\varepsilon_{t-j}$$

$$\delta_T^{-2T}\sum_{t=1}^{T} X_t(K)^2 \Longrightarrow K^2c^2Y^2. \tag{1.59}$$

In fact, using Cramér-Wold device it can be shown that the convergence in (1.58) and (1.59) is joint. Therefore, by the continuous mapping theorem

$$\widehat{\mathcal{VR}}(K) \sim \frac{\sum_{t=1}^{T} X_t(K)^2}{K\sum_{t=1}^{T} X_t^2} \xrightarrow{P} K,$$

as required. ■

Proof of Theorem 1.6. From (1.37) and (1.38) it is straightforward to see that

$$X_{t+i} = (t+i)\mu + \mu_0 + \sum_{j=1}^{t+i} \eta_j + \varepsilon_{t+i}$$

$$\mathrm{var}(X_{t+i}) = \Sigma_\varepsilon + \frac{t+i}{T}\Sigma_\eta \quad \text{and} \quad \mathrm{cov}(X_{t+i}, X_{t+k}) = \frac{t+i}{T}\Sigma_\eta$$

for $i, k = 0, 1, \ldots, K-1$ and $i < k$. Consequently we have

$$\mathrm{var}(X_t + X_{t+1} + \cdots + X_{t+K-1}) = K^2 \frac{\Sigma_\eta}{T}\left(t + \sum_{j=1}^{K-1}\left(1 - \frac{j}{K}\right)^2\right) + K\Sigma_\varepsilon$$

$$\begin{aligned}\mathcal{VR}_T(K) = & \left(\Sigma_\eta \frac{T+1}{2T} + \Sigma_\varepsilon\right)^{-1/2} \times \\ & \left[K\Sigma_\eta\left(\frac{T+1}{2T} + \frac{1}{T}\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)^2\right) + \Sigma_\varepsilon\right] \times \left(\Sigma_\eta \frac{T+1}{2T} + \Sigma_\varepsilon\right)^{-1/2}\end{aligned}$$

so that

$$\begin{aligned}\frac{1}{K}\lim_{T\to\infty} \mathcal{VR}_T(K) &= \left(\Sigma_\eta \frac{1}{2} + \Sigma_\varepsilon\right)^{-1/2}\left[\Sigma_\eta \frac{1}{2} + \frac{1}{K}\Sigma_\varepsilon\right]\left(\Sigma_\eta \frac{1}{2} + \Sigma_\varepsilon\right)^{-1/2} \\ &\longrightarrow \left(\Sigma_\eta \frac{1}{2} + \Sigma_\varepsilon\right)^{-1/2} \frac{1}{2}\Sigma_\eta \left(\Sigma_\eta \frac{1}{2} + \Sigma_\varepsilon\right)^{-1/2}\end{aligned}$$

as $K \to \infty$, completing the proof. ■

Proof of Theorem 1.7. It is straightforward to see that

$$\begin{aligned}\sqrt{T}\left[\frac{1}{d}\sum_{i=1}^{d}\left(\widehat{\mathcal{VR}d}_{+,ii}(K) - 1\right)\right] &= \sqrt{T}\cdot\frac{1}{d}\sum_{i=1}^{d}\left(2\sum_{j=1}^{K-1}\left(1-\frac{j}{K}\right)\frac{\widehat{\gamma}_{ii}(j)}{\widehat{\sigma}_{ii}}\right) \\ = \; & \frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{K-1}\frac{c_j}{\sigma_{ii}}\frac{1}{\sqrt{T}}\sum_{t=j+1}^{T}\widetilde{X}_{it}\widetilde{X}_{i,t-j} - \left[\max_{1\le i\le d}(\overline{X}_i - \mu_i)\right]\cdot\left\{\frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{K-1}\frac{c_j}{\sigma_{ii}}\frac{1}{\sqrt{T}}\sum_{t=1+j}^{T}\widetilde{X}_{it}\right\} \\ & - \left[\max_{1\le i\le d}(\overline{X}_i - \mu_i)\right]\cdot\left\{\frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{K-1}\frac{c_j}{\sigma_{ii}}\frac{1}{\sqrt{T}}\sum_{t=1+j}^{T}\widetilde{X}_{i,t-j}\right\} \\ & + \frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{K-1}\frac{c_j}{\sigma_{ii}}\frac{T-j-1}{\sqrt{T}}(\overline{X}_i - \mu_i)^2 + o_p(1) \\ = \; & \frac{1}{\sqrt{T}}\sum_{t=K}^{T}\left(\frac{1}{d}\sum_{i=1}^{d}\left[\sum_{j=1}^{K-1}\frac{c_j\widetilde{X}_{it}\widetilde{X}_{i,t-j}}{\sigma_{ii}}\right]\right) + o_p(1) \qquad (1.60)\end{aligned}$$

because by uncorrelatedness of $\widetilde{X}_t$

$$
\begin{aligned}
P\left(\left|\max_{1\le i\le d}(\overline{X}_i-\mu_i)\right|>\varepsilon\right) &\le \sum_{i=1}^{d}\frac{1}{\varepsilon^2}E\left(\frac{1}{T}\sum_{t=1}^{T}\widetilde{X}_{it}\right)^2 \\
&\le \frac{d}{\varepsilon^2}\frac{1}{T^2}\sum_{t=1}^{T}\max_{1\le i\le d}E\left(\widetilde{X}_{it}^2\right)=O\left(\frac{d}{T}\right),
\end{aligned}
$$

and by the moment-cumulant relationship formula and Assumption S we have

$$
\begin{aligned}
&P\left(\max_{1\le i\le d}\left|\frac{1}{T}\sum_{t=1}^{T}\widetilde{X}_{it}^2-E\big(\widetilde{X}_{it}^2\big)\right|>\varepsilon\right) \\
&\qquad\le \frac{1}{\varepsilon^2}\sum_{i=1}^{d}\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{Cov}\big(\widetilde{X}_{it}^2,\widetilde{X}_{is}^2\big) \\
&\qquad\le \frac{1}{\varepsilon^2}\sum_{i=1}^{d}\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\Big[2\big(\gamma_{ii}(t-s)\big)^2+\kappa_{iiii}(s,s,0,0)\Big]=O\left(\frac{d}{T}\right),
\end{aligned}
$$

from which it follows that $\max_i|\widehat{\sigma}_{ii}-\sigma_{ii}|=o_p(1)$. The $\{.\}$ terms can be easily shown to be bounded in probability using Chebyshev's inequality and uncorrelatedness of $\widetilde{X}_t$.

It now suffices to derive the limiting distribution of (1.60). We will only briefly sketch the proof as the main arguments closely follow those of proof of Theorem 1.2. Since the asymptotic variance

$$
\begin{aligned}
qd(\infty)^* &= \lim_{d\to\infty}\frac{1}{d^2}\tau^{\intercal}Qd(K)\tau \\
&= \lim_{d\to\infty}\frac{1}{d^2}\sum_{i=1}^{d}\sum_{r=1}^{d}\left[\sum_{j=1}^{K-1}\sum_{k=1}^{K-1}\frac{c_jc_k}{\sigma_{ii}\sigma_{rr}}E(\widetilde{X}_{it}\widetilde{X}_{rt}\widetilde{X}_{i,t-j}\widetilde{X}_{r,t-k})\right],
\end{aligned}
$$

where $\tau=\mathrm{vec}(I_d)$, is finite by Assumption Sd, we see that upon checking the required conditions the central limit theorem for martingale difference applies, yielding

$$
\frac{1}{\sqrt{T}}\sum_{t=K}^{T}\left(\frac{1}{d}\sum_{i=1}^{d}\left[\sum_{j=1}^{K-1}\frac{c_j\widetilde{X}_{it}\widetilde{X}_{i,t-j}}{\sigma_{ii}}\right]\right)\Longrightarrow N\big(0,qd(\infty)^*\big).
$$

Note that the conditional Lindeberg condition can be shown to hold by repeatedly using Minkowski's inequality and by Assumption A$'$. The remaining consistency result follows by the ergodic theorem and Assumption Sd, completing the proof. ■

Chapter 2

Nonparametric Estimation of Infinite Order Regressions[1]

Conditional expectations are crucially important in financial economics, with implications in many applications including asset returns predictability, market efficiency and risk management. One fundamental objective is to understand the risk/return trade-off summarized by the relationship between the *expected excess return* relative to the *conditional variance* of returns. Due to the latency of conditional expectations however, there has been no universal agreement upon what is the best way to measure these objects. Differences in the approaches to modelling and estimating the conditional mean and variance has led to disagreement on their measurement, and also, conflicting empirical evidence on their intertemporal relation. Theoretical asset pricing models do not generally restrict the shape of the risk premium or the dynamics of the risk return trade-off. For example, Backus and Gregory (1993) show that the shape of the relation between the risk premium and the conditional variance of returns is largely unrestricted with increasing, decreasing, flat, or nonmonotonic patterns all possible. Similar conclusions are drawn by studies such as Abel (1988), Gennotte and Marsh (1993), and Veronesi (2000). Nevertheless, most empirical studies adopt a simple linear specification.

There are some issues with the usual approaches. First, there is the risk of misspecification. For instance, some studies have relied on parametric or semiparametric assumptions such as the ARCH or stochastic volatility models, where some high degree of structure is imposed on the return generating process. Other studies have typically measured the conditional mean and conditional variance as projections onto some predetermined variables. These approaches cannot be entirely justified, since they are all

[1]Part of this chapter was written during my short stay at Institut de Mathématiques in Toulouse in November 2013. I thank Philippe Vieu for his kind hospitality and advice. Many helpful advice from John Aston, Xiaohong Chen, Jeroen Dalderop, Paul Doukhan, Jiti Gao, Shuyi Ge, Hayden Lee, Yuan Liao, Mikhail Lifshits, Richard Nickl, Alexei Onatski, Hashem Pesaran, Peter Phillips, Alessio Sancetta are gratefully acknowledged. Special thanks to Alexey Rudenko for providing an original Russian photocopy of Sytaya (1974), and to Hyungjin Lee for translating the paper.

necessarily prone to some degree of potential specification error, see Linton and Perron (2003) and Escanciano, Pardo-Fernández and Van Keilegom (2017) for further discussions. Nonparametric modelling can be an effective solution in this context. It is a well established practical tool for analyzing time series data; see for example Härdle (1990), Bosq (1996), or Fan and Yao (2003) for a comprehensive review. A major advantage of this approach is that the relationship between the explanatory variables under study, denoted by $X = (X_1, \ldots, X_d)^\intercal$, and the response, say Y, can be modelled without assuming any restrictive parametric or linear structures. Stone (1980, 1982) showed that the best achievable convergence rate (in minimax sense) is $n^{-\beta/(2\beta+d)}$, where β is a measure of smoothness and d is the dimension of the covariates.

Secondly, there may be potential bias due to the omission of necessary information. Choosing among a few conditioning variables introduces an element of arbitrariness into the econometric modelling of expectations. In particular, if information that investors consider important is neglected, then the corresponding estimates may be unreliable, Harvey (2001). Lettau and Ludvigson (2010) argued that contrasting conclusions on the intertemporal risk-return relation are largely due to the prevalent use of only small amount of conditional information in modelling the conditional mean and variance. Indeed, such practice greatly restricts the dynamics for the variance process and may result in poor estimates, especially when the volatility is highly persistent, Linton and Perron (2003), Giraitis et al. (2008). For example, Pagan and Hong (1990) estimated the conditional moments with nonparametric estimates of $\mathrm{E}(r_{mt} - r_{ft} | r_{m,t-1}, \ldots, r_{m,t-p})$ and $\mathrm{var}(r_{mt} - r_{ft} | r_{m,t-1}, \ldots, r_{m,t-p})$, where $r_{mt} - r_{ft}$ denotes the excess market return and $p = 1$ or 4. Having ended up with a negative risk-return relation using their estimates, they conjectured that the conclusion may have been affected by their use of only a small, finite number of conditioning variables. Noting the dependence of a GARCH process on the infinite past history of returns (with declining weights), they wrote: "*[A] nonparametric estimator of σ_t^2 appeals as a solution ..., although the fact that it operates with only a finite number of conditioning elements makes it unable to explicitly handle a GARCH type process. ... [O]ne might be able to establish consistency of the estimator [which deals with the infinite dependence]. As far as we are aware, however, there are no current theorems that would justify such a conjecture.*"

2.0.1 Overview of Results

This paper defines an estimation method that effectively addresses the aforementioned difficulties. We propose a Nadaraya-Watson type estimator that operates with an unrestricted number of conditioning variables. We derive large sample properties of the estimator in extensive detail, thereby providing an answer to the longstanding question in the quotation above. With a bandwidth sequence that shrinks the effects of long lags, the influence of all conditioning information is modelled in a natural and flexible way, and both issues of omitted information bias and specification error are effectively handled. It is worth noting that Harvey (2001) reported sensitivity of conditional expectations estimates on what type of conditioning variables are used in modelling the expectations. He showed with examples how several parametric/nonparametric estimates (and the estimated risk-return relationship) may vary according to the choice of different predetermined conditioning information. In this paper, we allow for various kinds of conditioning information. This is achieved by letting our model assumptions cover a wide range of static and dynamic regressions frameworks. The latter includes the autoregression framework as a special case.

Linton and Sancetta (2009) tackled this estimation problem of infinite order regression in the autoregression context. They established uniform almost sure consistency for stationary ergodic data but without rates. In the conclusion, they conjectured that the limiting distribution of nonparametric estimators could be established, and that the rate of convergence would be logarithmic. Under strict cross-sectional and temporal i.i.d. assumption, Mas (2012) derived a convergence rate that is consistent with our results in the particular case they considered.

We make several contributions. First, we establish some theorems which answer several open questions posed in the literature. Specifically, we show the pointwise consistency of our estimator under a set of mild regularity conditions. Further, we establish a central limit theorem for our estimator at a point under stronger conditions as well as for a feasibly studentized version of the estimator, thereby allowing pointwise inference to be conducted. Also, uniform consistency of the estimator is shown over a compact set of logarithmically increasing dimension. We prove that convergence rates depend on the smoothness of the regression function, the distribution of the marginal regressors and their dependence structure in a non-trivial way via the Lambert W function. We elaborate how each of those factors affects the rate of convergence, and show that the best possible rate is, nonetheless, of logarithmic order in all cases regardless of the smoothness of the regression function. This reflects the difficulty of capturing nonparametrically the effect of an infinite number of lags.

Second, using our estimation method we find some new empirical results. We reveal new evidence on the dynamics of risk-return relation and its link with the macroeconomy, and add supporting evidence for explaining some major puzzles in financial economics. To elaborate, applying our methods on the US stock market we find a positive risk-return relationship over the past 60 years overall − which is what asset pricing models generally postulate, e.g. Merton (1973). In particular, the relation turns out to be highly positive and strongly statistically significant in the recent 30 years period. Moreover, we also found that there has been a strong time variation and counter-cyclicity in risk aversion and in the conditional Sharpe ratio. The time series of estimated risk aversion tends to move in the opposite direction to the Federal Funds rate, a proxy for the business cycle, with the sample correlation being -0.5673. The quarterly Sharpe ratio is also strongly counter-cyclical, rising over most periods of recessions. By contrast, when a standard nonparametric method is employed instead, we noticed that these findings are not revealed, and different conclusions are reached. We believe our new empirical findings suggest an improvement in the econometric analysis that is attributable to allowing for extended flexibility and the inclusion of otherwise neglected information in our method.

2.0.2 Technical challenges and sketch proposals for remedy

One major hurdle we face in the infinite-dimensional setting is the non-existence of the usual notion of density $p(\cdot)$ for the regressor X. Since there is no σ-finite Lebesgue measure in infinite-dimensional spaces, the Lebesgue density (with respect to the infinite product of probability measures) of the regressor cannot be defined via the Radon-Nikodym theorem. Consequently, standard asymptotic arguments for kernel estimators are no longer valid, for example: Bochner's lemma whereby under suitable regularity conditions, for $j = 1, 2$

$$\begin{aligned} \frac{1}{h^d} \mathrm{E}\left[\mathcal{K}^j\left(\frac{x-X}{h}\right)\right] &= \int \mathcal{K}^j(u) p(x-uh)\ du \\ &\to p(x) \|\mathcal{K}\|_j^j \quad \text{as} \quad h \to 0 \end{aligned} \tag{2.1}$$

where $\mathcal{K}$ is a multivariate kernel (see Section 2.2.2 below). So classical limiting theories cannot be readily extended to our setting.

We propose to adopt and apply some ideas from the functional regression literature. There is a vast statistical literature on functional data (typical examples include curves and images, which are infinite-dimensional in nature). Ferraty and Vieu (2002) first studied the case where the regressor was function-valued. Masry (2005) provided a rigorous treatment of nonparametric regression with dependent functional data in which X lies in a general semi-metric space, establishing the central limit theorem. Mas (2012) derived the minimax rate of convergence for nonparametric estimation of the regression function with strictly independent and identically distributed covariates. Ferraty and Vieu (2006) detailed a number of extensions and gave an overview of nonparametric approaches in the functional statistics literature. Geenens (2011) gave an up-to-date accessible summary of the literature on nonparametric functional regression, and introduced the term *curse of infinite dimensionality*, which reflects the evident difficulties in nonparametric estimation of infinite-dimensional objects due to extreme data sparsity. In the finite dimensional case more smoothness can mitigate completely the slower rate of convergence caused by dimensionality, but in the infinite dimensional case, additional smoothness can only mildly improve the convergence rate of estimators. We discuss in the next section the difference between the functional data framework and our discrete time framework.

There is another potential problem that may arise specifically in the infinite dimensional setting. In the dynamic regression framework, the regressor vector X_t includes the infinite lags of a variable Z_t, say the response variable. Consequently, the class of mixing type assumptions, a popular notion of dependence in the econometrics literature, is generally not applicable. This is because measurable functions of X_t will depend upon the infinite time-lags of Z_t, and are not mixing in general, see e.g. Davidson (1994). Therefore, in order to establish asymptotic theories, an alternative set of dependence assumptions should be imposed on the data generating process. We defer further discussions to Section 2.2.1 below.

2.1 Some Preliminaries

Consider the regression model

$$Y = m(X) + \varepsilon, \tag{2.2}$$

where the regressor $X = (X_1, X_2, \ldots.)^\intercal$ is a random element taking values in some sequence space S, the response Y is a real-valued variable, and the stochastic error ε is such that $E(\varepsilon|X) = 0$ a.s. The objective is to estimate the Borel function

$$m(\cdot) = E(Y|X = \cdot) \tag{2.3}$$

based on n random samples observed from a strictly stationary data generating process $\{(Y_t, X_t) \in \mathbb{R} \times S\}_{t\in\mathbb{Z}}$ having some weak dependence structure (see Section 2.2.1 below).

This setting is related to the usual framework adopted for functional data, which has been widely studied by statisticians, see Ramsey and Silverman (2002), Aneiros, Bongiorno, Cao and Vieu (2017). Recently, successful attempts have been made to develop theories for nonparametric inference in the functional statistics literature; Ferraty and Romain (2010) gives a comprehensive review. A major issue in this field of research lies in extending the statistical theories applicable to $\mathbb{R}^d$ to function spaces. In this literature, attention is usually on smooth functions that are approximated and reconstructed from finely discretised grids on some compact interval. In contrast, the setup in our model (2.2) can be viewed as looking at a countable number of discrete observations. Such a difference is reflected by the fact that the observed data is taken to be a discrete process $X = (X_s)$ with unbounded $s \in \mathbb{Z}^+$ so that $S = \{f|f : \mathbb{N} \to \mathbb{R}\}$, rather than $X = (X(s))$ with $s \in [0, T]^k$ so that $S = \{f|f : [0, T]^k \subset \mathbb{R}^k \to \mathbb{R}\}$, e.g. curves if $k = 1$, images if $k \geq 2$. The discrete nature of our setting has several fundamental distinctive features that allow us to look further into many specific practical issues.

An immediate consequence of our framework is that the tuning parameter can be imposed on each and every dimension, allowing one to control the marginal influence of the regressors. For instance when it is sensible to postulate that the influence of distant covariates is getting monotonically downweighted, one may set the marginal bandwidths to increase in the lag horizon so as to impose higher amount of smoothing at distant lags. Depending on the nature of the regressor, S may be taken as the space of all infinite real sequences $\mathbb{R}^\infty := \prod_{j=1}^\infty \mathbb{R}_j$ formed by taking Cartesian products of the reals, or its various linear subspaces such as ℓ_∞, ℓ_p, c. We propose to take $S = \mathbb{R}^\infty$ so as to refrain from imposing any prior restrictions with regard to the choice of the regressor; for example, taking S to be the space of bounded sequences excludes the possibility of having regressors with infinite support (e.g. Gaussian process).

2.1.1 Dependence structure and leading examples

A distinctive characteristic of time series data is temporal dependence between observations. In the nonparametric time series literature, Rosenblatt (1956)'s α-mixing has been the *de facto* standard choice due to it being the weakest among the class of mixing-type asymptotic independence conditions. Roussas (1990) established pointwise and uniform consistency of the local constant estimator under this condition, respectively, while Fan and Masry (1992) established asymptotic normality. The α-mixing condition has also been widely used in the context of dependent functional observations, see for instance Ferraty et al. (2010), Masry (2005), and Delsol (2009).

Definition 2.1. *A stochastic process $\{Z_t\}_{t=1}^\infty$ defined on some probability space $(\Omega, \mathcal{F}, P)$ is called α-mixing (NB. 'jointly' α-mixing if Z_t is $\mathbb{R}^d$-valued, with $d \in (1, \infty]$) if*

$$\alpha(r) := \sup_{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{r+t}^\infty} |P(A \cap B) - P(A)P(B)|$$

is asymptotically zero as $r \to \infty$, where $\mathcal{F}_a^b$ is the σ-algebra generated by $\{Z_s; a \leq s \leq b\}$. In particular, we say the process is algebraically (respectively exponentially) α-mixing with rate k if there exists some $c, k > 0$ such that $\alpha(r) \leq cr^{-k}$ (respectively if there exists some $\gamma, \varsigma > 0$ such that $\alpha(r) \leq \exp(-\varsigma r^\gamma)$).

The popularity of the α-mixing condition (note the modifier α- will occasionally be omitted if no confusion is likely) in the literature stems from the fact that it is easy to work with, see e.g. Doukhan (1994) or Rio (2000) for a comprehensive survey. However, there are several limitations that have been pointed out in the literature. First, it is a rather strong technical condition that is hard to verify in practice. Second, some basic processes are not mixing. e.g. AR(1) with Bernoulli innovations, Andrews (1984).

We turn to our setting. In the static regression case it is appropriate to assume the mixing condition, but in the dynamic case this condition is not generally applicable as we now explain. Recall that the object of estimation is the conditional mean $E(Y_t|\mathcal{F})$, see (2.3), where the information set $\mathcal{F}$ is determined by the nature of the conditioning variables. There are two leading cases: the first case is the static regression where the information set is taken to mean $\sigma(X_{jt}; j = 1, 2, \ldots)$, the σ-algebra generated by the exogenous marginal regressors. The second case is the autoregression, where $X_{tj} = Y_{t-j}$ for all j, in which case $\mathcal{F} = \mathcal{F}_{t-1}$ represents $\sigma(Y_s; s \leq t-1)$, the σ-algebra generated by the sequence of lags of the response $(Y_s)_{s \leq t-1}$. In fact, as for the latter framework we may consider a more general setup, i.e. a dynamic regression, where the information set is taken to be $\mathcal{F} = \sigma(X_{js}, Y_s;\ s \leq t-1)$ for some j. Details are formally given in Assumptions A below.

In the static regression case the usual joint α-mixing condition can be assumed on the sample data $\{Y_t, X_t\}$ as is usually done; since marginal regressors are observed at the same time t: $X_t = (X_{1t}, X_{2t}, \ldots)^\intercal$, assuming joint dependence does not require additional adjustments. Indeed, it can be easily shown that joint mixing implies both marginal component processes and any measurable function thereof are mixing.[2] In this chapter, we do not necessarily require independence between component processes $\{X_{jt}\}$, $j = 1, 2, \ldots$; later we specify to what extent some dependence can be allowed (see Assumption C2). It will turn out that the requirement is mild and allows sufficient generality in application.

Moving on to the dynamic regression setting, since the regressors are taken to be the lags of the response and/or a covariate, measurable functions of X_t depend on infinite time-lags and hence are *not* necessarily mixing.[3] Therefore an alternative set of dependence conditions is necessary to establish asymptotic theories for the second framework. We adopt the notion of near epoch dependence due to Ibragimov (1962) for the dynamic regression setting and deal with two leading cases separately.

Definition 2.2. *A stochastic process $\{Z_t\}_{t=1}^{\infty}$ defined on some probability space $(\Omega, \mathcal{F}, P)$ is called near-epoch dependent or stable in L_2 with respect to a strictly stationary α-mixing process $\{\eta_t\}$ if the stability coefficients $v_2(r) := E|Z_t - Z_{t,(r)}|^2$ is asymptotically zero as $r \to \infty$, where $Z_{t,(r)} = \Psi_r(\eta_t, \ldots, \eta_{t-r+1})$ for some Borel function $\Psi_r : \mathbb{R}^r \to \mathbb{R}$.*

A process that is *near epoch* dependent on a mixing sequence is influenced primarily by the "recent past" of the sequence and hence asymptotically resembles its dependence structure; see e.g. Billingsley (1968), Davidson (1994), or Lu (2001) for details. Andrews (1995) established uniform consistency of kernel regression estimators under near epoch dependence conditions. Following the usual convention, e.g. Bierens (1983), we shall

[2] The converse is not necessarily true unless the marginal processes are independent to each other, see Bradley (2005, Section 5).

[3] Except for some very special cases; Davidson (1994, Theorem 14.9) gives a set of technical conditions under which a process with infinite (linear) temporal dependence is α-mixing.

take $\Psi_r(\eta_t, \ldots, \eta_{t-r+1}) \equiv E(Z_t|\eta_t, \ldots, \eta_{t-r+1})$. In Section 2.3 it will be shown that under suitable conditions similar asymptotic theories can be derived for both static and dynamic regression frameworks.

2.1.2 Local Weighting

In this section we fix the notions of local weighting and the measure of closeness between the data objects. Let $K : [0, \infty) \to [0, \infty) =: \mathbb{R}_+$ be a univariate density function and for an element u of a normed sequence space, let

$$\mathcal{K}(u) := K(\|u\|). \tag{2.4}$$

In our setting the properties of K are crucially important. We now group the kernel functions into three subcategories depending on how they are generated. The first two, referred to as Type-I and Type-II kernels in Ferraty and Vieu (2006) generalize the usual 'window' kernels and monotonically decreasing kernels in finite dimension, respectively. Both types of kernels are continuous on a compact support $[0, \lambda]$.

DEFINITION 2.3. *A function $K : [0, \infty) \to [0, \infty)$ is called a kernel of type$-I$ if it integrates to 1, and if there exist real constants C_1, C_2 (with $0 < C_1 < C_2$) for which*

$$C_1 1_{[0,\lambda]}(u) \le K(u) \le C_2 1_{[0,\lambda]}(u), \tag{2.5}$$

where λ is some fixed positive real number. A function $K : [0, \infty) \to [0, \infty)$ is called a kernel of type$-II$ if it satisfies (2.5) with $C_1 \equiv 0$, and is continuous on $[0, \lambda]$ and differentiable on $(0, \lambda)$ with the derivative K' that satisfies

$$C_3 \le K'(u) \le C_4$$

for some real constants C_3, C_4 such that $-\infty < C_3 < C_4 < 0$.

The definition above suggests that the uniform kernel on $[0, \lambda]$ is a type-I kernel, and the Epanechnikov, Biweight and Bartlett kernels belong to the class of Type-II kernels. Some of those with semi-infinite support, for example (one-sided) Gaussian, are covered by the last group, which we will call the Type-III kernels.

DEFINITION 2.4. *A function $K : [0, \infty) \to [0, \infty)$ is a kernel of type$-III$ if it integrates to 1, and if it is of exponential type; that is, $K(r) \propto \exp(Cr^\beta)$ for some β and C.*

2.1.3 Small deviations

The *small ball (or small deviation) probability* plays a crucial role in establishing the asymptotic theory. Let S^* be a sequence space equipped with some norm $\|.\|$; then the small ball probability of an S^*-valued random element Z is a function defined as

$$\varphi_z(h) := P\left(\|z - Z\| \leq h\right), \tag{2.6}$$

where $h \in \mathbb{R}_+$. The probability is called *centered* if $z = 0$ (in which case we write $\varphi(h)$ and *shifted* (with respect to some fixed point $z \in S^* \backslash \{0\}$) if otherwise. The relation between the two quantities cannot be explicitly specified in general, and will be given in terms of the Radon-Nikodym derivative (See Assumption D1 below).

The name *small ball* stems from the fact that we are interested in the asymptotic behaviour of $\varphi_z(h)$ as h tends to zero. The function can be thought of as a measure for how much the observations are densely *packed* or *concentrated* around the fixed point z with respect to the associated norm and the reference distance h. From the definition it is straightforward to see that $\varphi_z(h) \to 0$ as $h \to 0$, and that $n\varphi_z(h)$ is an approximate count of the number of observations whose influence is taken into account in the smoothing procedure. When Z is a continuous random vector of fixed dimension d with density $p(\cdot) > 0$, it can be readily shown that the shifted small ball probability (with respect to the usual Euclidean norm) is given by

$$\varphi_z(h) = V_d h^d p(z) = O(h^d), \tag{2.7}$$

where $V_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of the d-dimensional unit sphere.

However, when Z takes values in an infinite-dimensional normed space, it is difficult to specify the exact form of the small ball probability, and its behaviour varies depending heavily on the nature of the associated space and its topological structure. Due to the non-equivalence of norms in infinite dimensional spaces, it is intuitively clear that the "speed" at which $\varphi_z(h)$ converges to zero is affected by the choice of the norm $\|.\|$. Nonetheless, because the data becomes sparser as the dimension increases, it is expected that the small ball probability would be lower and tend to zero faster in general in infinite dimensional spaces.

One possible example of S^* is $(\ell_r, \|.\|_r)$, the space of r-th power summable sequence equipped with the ℓ_r-norm; the centred small ball behaviour of sums of weighted i.i.d. random variables is widely studied in the literature, see for example Borovkov and Ruzankin (2008) and references therein. In this work here, we will focus our main attention on the case of $r = 2$ (and take $\|.\|$ to mean $\|.\|_2$ unless specified otherwise). Nevertheless, we note that the results derived in this chapter can be extended to the case of $r > 2$ as long as the regularity conditions are adjusted appropriately.

Writing the expected value of the kernel in terms of the small ball probability

$$E\mathcal{K}\left(\frac{z-Z}{h}\right) = EK\left(\frac{\|z-Z\|}{h}\right) = \int K(u)\ dP_{\|z-Z\|/h}(u) = \int K(u)\ d\varphi_z(uh), \quad (2.8)$$

we are able to bypass the difficulties mentioned in the introduction, and to establish the convergence of the integrals without explicitly requiring the existence of the Lebesgue density.

Lemma 2.1. Ferraty and Vieu (2006, Lemma 4.3 & 4.4). *Suppose $\|.\|$ is some semi-norm defined on a function space. If K is type-I, then it satisfies*

$$C_1^j \leq \frac{1}{\varphi_z(h\lambda)} \int_0^\lambda K^j(v)\ d\varphi_z(vh) \leq C_2^j, \quad j = 1, 2 \quad (2.9)$$

where $C_1, C_2 > 0$ are as defined in Definition 2.3. When the kernel K is type-II, if

$$\exists\ \varepsilon_0 > 0, C_5 > 0 \ s.t.\ \forall \varepsilon < \varepsilon_0, \quad \int_0^\varepsilon \varphi_x(u)du > C_5 \varepsilon \varphi_x(\varepsilon) \quad (2.10)$$

then we have

$$C_6^j \leq \frac{1}{\varphi_z(h\lambda)} \int_0^\lambda K^j(v)\ d\varphi_z(vh) \leq C_7^j, \quad j = 1, 2 \quad (2.11)$$

where the constants $C_6 = -C_5 C_4$ and $C_7 = \sup_{s \in [0,\lambda]} K(s)$ are strictly positive.

Under the regularity conditions of Lemma 2.1, (2.9) and (2.11) hold for every $h > 0$, so it follows that for any kernels of type-I and II:

Corollary 2.1. *If the kernel K is either type-I or type-II, then for $j = 1, 2$ we have*

$$\frac{1}{\varphi_z(h\lambda)} E\left[\mathcal{K}^j\left(\frac{z-Z}{h}\right)\right] \longrightarrow \xi_j \quad as \quad h \to 0^+, \quad (2.12)$$

where ξ_1 and ξ_2 are some strictly positive real constants.

This result can be seen as an infinite-dimensional analogue of Bochner's lemma (2.1): i.e., for $Z \in \mathbb{R}^d$, $h^{-d} E\mathcal{K}((z-Z)/h) \to p(z) > 0$. It is obvious that ξ_j is bounded below and above by C_1^j and C_2^j, respectively (or C_6^j and C_7^j depending on the choice of the kernel). With specific choices of kernels and regressors we may be able to specify the exact values of the constants in some certain cases. For example, it is straightforward to see that $\xi_1 = 1/\lambda$ and $\xi_2 = 1/\lambda^2$ when K is uniform kernel.

Remarks. (i) Lemma 2.1 reveals the importance of condition (2.10) in constructing the asymptotics when the kernel is of type-II. Whereas the condition is widely assumed in the functional statistics literature for that reason, Azais and Fort (2013) proved that it necessarily restricts the variable Z to be of finite dimension. In other words, whenever (2.10) is valid, the topology that governs the concentration properties of Z accounts effectively only for finite dimension. An example (cf. Section 13.3.3 of Ferraty and Vieu (2006)) includes the case where Z is associated with the semi-norm $\|y\| := (y_1, \ldots, y_p, 0, 0, \ldots)$ for some positive integer $p < \infty$, where $y \in \mathbb{R}^\infty$. This severely restricts the applicability of our work. Because no sufficient condition other than (2.10) for (2.11) is known in the literature, we shall not cosider the case of Type-II kernels in this work.

(ii) A natural question one may then ask is whether (2.12) would hold for kernels with semi-infinite support such as the Type-III kernels. In the finite $\mathbb{R}^d$-framework, it is well known that a set of assumptions including $\|u\|^d K(u) \to 0$ as $u \to \infty$ is sufficient for showing (2.1), see for instance Parzen (1962, Theorem 1A) and Pagan and Ullah (1999, Lemma 1). However, in the infinite-dimensional setting the answer is negative in most usual cases where the kernel is of exponential type (e.g. Gaussian kernel). Whereas the lower bound of the limit can be easily constructed via Chebyshev's inequality: with reference to Definition 2.4, writing $V = \|z - Z\|^\beta$, $\delta = h^\beta$ and letting c_δ be some function of δ we have

$$(0 <) \exp(-c_\delta \delta) \leq [P(V \leq \delta)]^{-1} E \exp(-c_\delta V). \tag{2.13}$$

So the upper bound may not exist, and the rate at which the small ball probability decays to zero may dominate the speed at which the integral (2.8) converges to zero. This claim cannot be formally verified for all general cases because (as aforementioned) there is no unified result for the asymptotic behaviour of small deviations available. Nevertheless, the idea can be sketched in the common case where the asymptotics of the distribution function (i.e. small deviation) is of exponential order: $P(V \leq \delta) \sim \exp(-C\delta^{-\theta})$ as $\delta \to 0$ for some constants C and $\theta > 0$. By de Bruijn's exponential Tauberian theorem (see Bingham et al. (1987), Li (2012)), a necessary and sufficient condition for such a case is the following limiting behaviour of the Laplace transform near infinity:

$$E[\exp(-c_\delta V)] \sim \exp\Big(- C' \cdot c_\delta^{\theta/(1+\theta)} \Big) \quad \text{as} \quad c_\delta \to \infty$$

for some constant $C' > 0$. With $V = \|z - Z\|^2$, $\delta = h^2$, $c_\delta = 2^{-1} h^{-2}$ (which corresponds to the case of the Gaussian kernel) the difference in the order of convergence suggests that the right hand side of (2.13) is unbounded, and that the limit (2.12) diverges.

Due to the reasons above we will confine our attention to Type-I kernels only here in this work.

2.1.4 Bandwidth Matrix and covariates

We aim to estimate the regression operator at a point $x \in \mathbb{R}^\infty$ with an $\mathbb{R}^\infty$-dimensional regressor $X = (X_1, X_2, \ldots)^\intercal$. Let $H := \operatorname{diag}(\underline{h}) = \operatorname{diag}(h_1, h_2, \ldots) \in \mathbb{R}^{\infty\times\infty}$ be the bandwidth matrix. We require that a norm $\|.\|$ can be admitted to the *weighted regressor* values and the *weighted point*, and for this the bandwidth sequence must be chosen appropriately. In particular, we let

$$H = hD = h \times \operatorname{diag}(\phi_1, \phi_2, \ldots), \tag{2.14}$$

where $D \in \mathbb{R}^{\infty\times\infty}$ and $h \in \mathbb{R}$. By Kolmogorov's three-series theorem, the sequence of weighted regressors $\{\phi_j^{-1} X_j\}$ is square summable, with probability one, provided that the marginal regressors X_j' are independent with finite variance and satisfy

$$\sum_{j=0}^{\infty} E \min\left\{1, \phi_j^{-2} X_j^2\right\} < \infty, \tag{2.15}$$

so that $(\phi_1^{-1} X_1, \phi_2^{-1} X_2, \ldots)^\intercal =: Z$ is $(\ell_2, \|.\|_2)$-valued. In the autoregressive framework, ϕ_j^{-1} can be interpreted as a weight sequence that represents the "relative influence" of the marginal regressors, which diminishes as lags get further apart.

For this purpose we assume from now on that *the bandwidth-weighted* X and x (i.e. Z and $z := (\phi_1^{-1} x_1, \phi_2^{-1} x_2, \ldots)^\intercal$, respectively) are ℓ_2-valued[4] and normed with $\|.\| = \|.\|_2$. Consequently, (with an abuse of notation) we can extend the usual definition of shifted small deviation to account for the generalized support $[0, \lambda]$ and bandwidth vector $\underline{h} = (h_1, h_2, \ldots)^\intercal$:

$$\begin{aligned} \varphi_x(\underline{h}\lambda) &:= P\left(\|H^{-1}(x - X_t)\| \le \lambda\right) \\ &= P\left(\|D^{-1}(x - X_t)\| \le h\lambda\right). \end{aligned} \tag{2.16}$$

Equivalently, $\varphi_x(\underline{h}\lambda) = P(X_t \in \mathcal{E}(x, \underline{h}\lambda))$, where $\mathcal{E}$ is the infinite-dimensional hyperellipsoid centred at $x \in \mathbb{R}^\infty$, and λ is as defined in Section 2.2. Clearly, $\varphi_x(\underline{h}\lambda) = \varphi_z(h\lambda)$. For later reference, we also define the joint small ball probability of the regressor vectors observed at different times t and s as the joint distribution

$$\psi_x(\underline{h}\lambda; t, s) := P\big((X_t, X_s) \in \mathcal{E}(x, \lambda\underline{h}) \times \mathcal{E}(x, \lambda\underline{h})\big). \tag{2.17}$$

[4] This gives a mild restriction on the range of possible points at which the estimation is made; i.e. $x \in \mathbb{R}^\infty$ is such that $\sum_j j^{-2p} x_j^2 < \infty$.

2.2 The Estimator

We observe a sample $\{Y_t, X_t\}_{t=1}^n$ with $Y_t \in \mathbb{R}$ and $X_t \in \mathbb{R}^\infty$. With these data, we propose to estimate $m(x) = E(Y|X = x)$, $x \in \mathbb{R}^\infty$ with the following local constant type estimator:

$$\widehat{m}(x) := \frac{\sum_{t=1}^n \mathcal{K}\Big(H^{-1}(x - X_t)\Big)Y_t}{\sum_{t=1}^n \mathcal{K}\Big(H^{-1}(x - X_t)\Big)} \equiv \frac{\sum_{t=1}^n K\Big(\|H^{-1}(x - X_t)\|\Big)Y_t}{\sum_{t=1}^n K\Big(\|H^{-1}(x - X_t)\|\Big)}. \tag{2.18}$$

In practice, in the autoregression case we essentially observe only $\{Y_1, Y_2, \ldots, Y_n\}$ rather than the full infinity, so further lags can be regarded as zeros. Similarly, in the static case, when X_t is in $\mathbb{R}^\tau$ for large τ we can identify this with $X_t = (X_{1t}, X_{2t}, \ldots, X_{\tau t}, 0, 0, \ldots)$ $\in \mathbb{R}^\infty$. So for practical applications, one may for example employ a truncation argument on the regressor as will be done in Section 2.4.4 - albeit with a different purpose) and let the effective dimension τ of the regressor X_t to increase in n in the theoretical analysis.

The estimator can be viewed as an infinite-dimensional generalization of the standard multivariate local linear estimator, and is a special case of the one in Ferraty and Vieu (2002), Masry (2005) and references therein for functional data. In the following section we will examine some asymptotic properties of the estimator.

2.3 Asymptotic Properties

In this section we introduce the main results of this chapter. We derive some large sample asymptotics of the proposed estimator (2.18). We establish consistency in both the pointwise and uniform sense, and also the asymptotic normality. All proofs are detailed in the appendix.

Consider two different cases: (1) the static regression and (2) the dynamic regression. Below we specify two sets of temporal dependence conditions, either of which will be assumed on the data generating process of the sample observations. Assumption A1 corresponds to the static regression case where we have exogenous regressors that are jointly observed in time in a weakly dependent manner. No restriction is needed as regards the dependence structure between the marginal regressors, although certain additional conditions can be potentially imposed at the later stage (see Assumptions C below). The second option A2 concerns with the dynamic regression framework. In this case, the notion of near epoch dependence is adopted to describe the dependence structure of the processes defined as functions of the response variables. The assumptions below suggest that there is a trade-off between the degree of mixing and the possible order of moments we allow on the response variable, i.e. $2 + \delta$.

Assumptions A

A1. *The marginal regressors* $X_{1t}, X_{2t}, \ldots$ *are exogenous variables, and the sample data* $\{Y_t, X_t\}_{t=1}^n = \{Y_t, (X_{1t}, X_{2t}, \ldots)\}_{t=1}^n$ *is stationary and jointly arithmetically* α*-mixing with rate* $k \geq 2(\delta+2)/\delta$*, where* δ *is as defined in Assumption B4 below.*

A2. *Each regressor is either a lag of the response variable* Y_t *or of a covariate* V_t*, i.e.* $X_{jt} = Y_{t-j}$ *or* $X_{jt} = V_{t-j}$*,* $j \in \mathbb{N}$*, and* $\{Y_t, V_t\}_{t=1}^n$ *is stationary and arithmetically* α*-mixing with rate* $k \geq 2(\delta+2)/\delta$*. Also, the process* $K_t := K(\|H^{-1}(x - X_t)\|)$ *is near epoch dependent on* (Y_t, V_t)*, and there exists some* $r = r_n \to \infty$ *such that the rate of stability for* K_t *denoted* $v_2(r_n) = v_2(r)$ *satisfies*

$$v_2(r)^{1/2}[\varphi_x(\underline{h}\lambda)]^{-(2\delta+3)/(2\delta+2)}n^{1/(2(\delta+1))} \to 0 \qquad \text{as} \quad n \to \infty. \tag{2.19}$$

Remark. Our model under Assumption A2 can be viewed as a generalization of the NAARX model in Chen and Tsay (1993). The framework nests both the fully autoregressive framework in which $X_{jt} = Y_{t-j}$ $\forall j$, and the case where the regressor vector consists only of the lags of a covariate V_t. Doukhan and Wintenberger (2008) studied the autoregressive models of order $d = \infty$ under a notion of weak dependence, and showed the existence of a stationary solution. This result was further studied in Wu (2011).

Some examples of Near Epoch Dependent processes and detailed discussions can be found in e.g. Davidson (1994).

2.3.1 Pointwise consistency

Pointwise consistency of the local constant estimator was first studied by Watson (1964) and Nadaraya (1964) for i.i.d data with $d = 1$. Their result was extended to the multivariate case (finite dimension) by Greblicki and Krzyzak (1980) and Devroye (1981). Robinson (1983) and Bierens (1983) were amongst the earliest papers that worked on consistency of the estimator with dependent observations (both static regression and autoregression were allowed in their frameworks), followed by Roussas (1989), Fan (1990), and Phillips and Park (1998) to name a few out of numerous papers. The case of the functional regressor was first studied by Ferraty and Vieu (2002).

In this section we establish the pointwise weak consistency of the estimator (2.18) with dependent data satisfying either A1 or A2. A set of assumptions required for the theory is now introduced, and some introductory arguments are briefly sketched.

Assumptions B

B1. *The regression operator* $m : \mathbb{R}^\infty \to \mathbb{R}$ *is continuous in some neighbourhood of* x

B2. *The marginal bandwidths satisfy* $h_j = h_{j,n} \to 0$ *as* $n \to \infty$ *for all* $j = 1, 2, \ldots,$ *where* $\mathrm{diag}(h_1, h_2, \ldots) = \mathrm{diag}(\underline{h}) = H$ *is the bandwidth matrix, and the small ball probability obeys* $n\varphi_x(\underline{h}\lambda) \to \infty$ *for every point* $x \in \mathbb{R}^\infty$*, where* $\varphi_x(\underline{h}\lambda) := P(\|H^{-1}(x - X)\| \le \lambda) \to 0$ *as* $n \to \infty$.

B3. *The kernel* K *is type*$-I$

B4. *The response* Y_t *satisfies* $E\big(|Y_t|^{2+\delta}\big) \le C < \infty$ for some $C, \delta > 0$.

B5. *The joint small ball probability (2.17) satisfies* $\psi_x(\underline{h}\lambda; i, j) \le C\varphi_x(\lambda\underline{h})^2,\ \forall i \ne j$.

B6. *The conditional expectation* $E\big(|Y_tY_s||X_t, X_s\big) \le C < \infty$ *for all* t, s.

Remark. The continuity assumption B1 is necessary for asymptotic unbiasedness of the estimator. It will be shown that the estimator is unbiased at every point of continuity, and that the rate of convergence for the bias term can be specified upon imposing further smoothness condition on the regression operator, see later. Assumption B2 can be thought of as an extension of the usual bandwidth conditions that are assumed in finite-dimensional nonparametric literature, cf. (2.7). As discussed before, $n\varphi_x(\underline{h}\lambda)$ can be understood as an approximate number of observations that are "close enough" to x. Therefore, it is sensible to postulate that $n\varphi_x(\underline{h}\lambda) \to \infty$ as $n \to \infty$, meaning that the point x is visited many times by the sample of data as the size of the sample grows to infinity. This is in line with the usual assumption that $nh^d \to \infty$ when $X \in \mathbb{R}^d$, in which case the small ball probability is given by $\varphi_x(h) \propto h^d p_X(x)$ as noted in (2.7). Conditions B5 and B6 are imposed to control the asymptotics of the covariance terms. The validity of condition B5 can be easily seen in the $\mathbb{R}^d$ frameworks; for relevant discussions, see Ferraty and Vieu (2006, Remark 11.2).

To sketch the idea, we write $K_t := K(\|H^{-1}(x - X_t)\|)$ for the sake of simplicity of presentation (note its dependence upon X_t), and express the estimator (2.18) as

$$\widehat{m}(x) := \frac{\sum_{t=1}^n K\big(\|H^{-1}(x - X_t)\|\big)Y_t}{\sum_{t=1}^n K\big(\|H^{-1}(x - X_t)\|\big)} = \frac{\frac{1}{n}\sum_{t=1}^n \frac{K_t}{EK_1} Y_t}{\frac{1}{n}\sum_{i=1}^n \frac{K_t}{EK_1}} = \frac{\widehat{m}_2(x)}{\widehat{m}_1(x)}. \tag{2.20}$$

We then employ the following decomposition:

$$\begin{aligned}\widehat{m}(x) - m(x) &= \frac{\widehat{m}_2(x)}{\widehat{m}_1(x)} - m(x) = \frac{\widehat{m}_2(x) - m(x)\widehat{m}_1(x)}{\widehat{m}_1(x)} \\ &= \frac{E\widehat{m}_2(x) - m(x)E\widehat{m}_1(x)}{\widehat{m}_1(x)} + \frac{[\widehat{m}_2(x) - E\widehat{m}_2(x)] - m(x)[\widehat{m}_1(x) - E\widehat{m}_1(x)]}{\widehat{m}_1(x)},\end{aligned} \tag{2.21}$$

where clearly $E\widehat{m}_1(x) = 1$. Below we show consistency by proving that the 'bias part' $E\widehat{m}_2(x) - m(x)$ and the 'variance part' $[\widehat{m}_2(x) - E\widehat{m}_2(x)] - m(x)[\widehat{m}_1(x) - 1]$ are both negligible in large samples. As for the latter term, it suffices to show the mean squared convergence of $\widehat{m}_2(x) - E\widehat{m}_2(x)$ to zero because $\widehat{m}_1(x) \to^P 1$ then readily follows.

Theorem 2.1. *Suppose that Assumptions B1-B5 hold. Then the estimator (2.18) with sample observations $\{Y_t, X_t^\intercal\}_{t=1}^n$ satisfying either A1 or A2 is weakly consistent for the regression operator $m(x) = E(Y|X = x)$. That is, as $n \to \infty$*

$$\widehat{m}(x) \xrightarrow{P} m(x). \tag{2.22}$$

In the following section, we present the rates of convergence and asymptotic normality under additional regularity conditions.

2.3.2 Asymptotic Normality

Earlier studies on the limiting distribution of the standard Nadaraya-Watson estimator can be traced back to Schuster (1972) and Bierens (1987), where the case of univariate and multivariate regressors was considered, respectively. The case of dependent samples was studied in Robinson (1983), Bierens (1983), Masry and Fan (1997), and by many others under various model setups and different regularity conditions. Masry (2005, Theorem 4) and Delsol (2009) established general distribution theories for Nadaraya-Watson type estimators in a semi-metric space. Our results are different from those in two respects. First, the difference of our framework from the functional literature discussed in the beginning of Section 2.2 gives us further flexibility, without which the analysis cannot be done to meet our specific needs. Second, whereas the final results of many existing papers were given in terms of abstract functions, our results are presented with an explicit rate of convergence, allowing practical applications.

The primary objective of this section is to outline the main theory and some interesting consequences thereof. Both cases of the independent marginals (i.e., when the marginal regressors X_{jt} are independent and identically distributed for each fixed t) and also a dependent framework are allowed. Specifically, we introduce how independence restriction can possibly be moderated to allow for some mild dependence structure. In particular, Assumption C2 below specifies the extent to which certain cross-sectional dependence can be allowed on the marginal regressors in our theory while still allowing for specification of the exact form of the convergence rate of the estimator.

Assumptions C. *For every fixed t, the real-valued stochastic process formed by the marginal regressors $\{X_{jt}\}_{j=1}^\infty$ is either:*

C1. *independent and identically distributed over j with $EX_{jt}^4 \le C < \infty$ $\forall j$*, or

C2. *stationary (over j), and admits a moving average representation:*

$$X_{jt} = \sum_{u=-\infty}^{\infty} a_u \epsilon_{j-u,t}, \tag{2.23}$$

where a_u is a square summable sequence, and $\{\epsilon_{jt}\}_j$ is an independent and identically distributed standard Gaussian sequence.

Remark. In either case the marginal regressors are required to be identically distributed over j; an additional distributional assumption will be imposed in D2 below. Nonetheless, the possible degree of dependence allowed in C2 is very mild and general, since an equivalent condition of having the representation for a Gaussian process is simply the existence of the spectral density. Note that (2.23) includes the causal (one-sided) MA representation as a special case. If a stationary stochastic process $\{X_{jt}\}_j$ is α-mixing (over j), then it always has such a representation (i.e. $a_u = 0,\ \forall u < 0$) provided it is Gaussian. This is because any α-mixing process is regular[5] by definition, so is linearly regular when it is Gaussian, and hence (with stationarity) admits the Wold decomposition with independent Gaussian innovations by Corollary 17.3.1 of Ibragimov and Linnik (1971).

Note that each C1 and C2 is consistent with the case allowed in Assumption A1 and A2, respectively (because in the latter case the process $\{X_{jt}\}_j$ consists of temporal lags of the response variable and/or a covariate which form a mixing process by Assumption A2), although the dependence structure specified in C2 can be allowed also for the static case (i.e. A1). This suggests that there is absolutely no need to assume independence between marginal regressors in our model (2.2) under Gaussianity, and hence a wide flexibility is allowed in terms of the model setup. In particular, the convergence rates of our estimator will be shown to be of the same order under both C1 and C2. Lastly, the requirement of finite 4th moment is imposed to ensure that the squared marginal regressors have finite second moments due to the reasons to be clarified below; obviously, when a lag of the response is included in the dynamic regression framework (A2), this forces $\delta \geq 2$ in B4.

We now introduce some main assumptions needed for distributional theories.

(1) The 'bias component'

The first part concerns with the asymptotic 'bias', where Assumptions B is strengthened by imposing additional smoothness conditions and suitable bandwidth adjustments. They belong to a set of sufficient conditions under which the exact upper bound of the asymptotic bias can be specified. Note that alternatively, a Fréchet differentiability-type

[5]In the sense of Ibragimov and Linnik (1971) and Davidson (1994, Part III)

condition may be imposed.

Further Assumptions B

B7. *The regression operator* $m : \mathbb{R}^\infty \to \mathbb{R}$ *satisfies*

$$\big|m(x) - m(x')\big| \le \sum_{j=1}^{\infty} c_j \big|x_j - x_j'\big|^\beta \tag{2.24}$$

for every $x, x' \in \mathbb{R}^\infty$*, and some constant* $\beta \in (0,1]$*, where* $\{c_j\}$ *is some sequence of real constants that satisfies* $\sum_{j=1}^\infty c_j \le 1$.

B8. *The marginal bandwidths satisfy* $h_j = \phi_j \cdot h$ *for some positive real numbers* ϕ_j*, where* $h = h_n \to 0$ *as* $n \to \infty$*. We suppose that* ϕ_j *satisfy* $\sum_{j=1}^\infty \phi_j^{-2} < \infty$ *and* $\sum_{j=1}^\infty c_j \phi_j^\beta < \infty$.

Remark. These additional assumptions help us specify and regulate the bias component. Assumption B8 extends the previous bandwidth condition B2. Obviously, it is consistent with (yet stronger than) what was previously assumed in B2, since $h \to 0$ implies the coordinate-wise convergence of each marginal bandwidths. With this condition one is able to write the asymptotic bias expression and the order of the bias-variance balancing bandwidth in terms of the common factor h. It is possible to dispense with this condition at the cost of imposing minor modifications in B7; the asymptotic bias will then be written in terms of the infinite sum of a weighted marginal bandwidth h_j, whose convergence needs to be ensured. To understand the asymptotic behaviour of the variance component, a further increment condition will be imposed on the sequence of marginal coefficients ϕ_j in Assumption D later. We remark that at this point such an assumption is not necessary as we are not concerned with the variance term.

Assumption B7 replaces and strengthens B1, and can be thought of as a variant of Hölder-type continuity; the case of $c_j = 2^{-j}$ and $\beta = 1$ is implied by the Lipschitz condition. Another example of c_j includes $\exp(-j)$. Indeed, under B7 the regression operator becomes a contraction mapping, and the contribution from each marginal dimension decreases in lag or index. This ensures summability of the bias of the estimator and allows one to specify its order of convergence rate, cf. (2.30) below.

In the context of autoregression where $X_j \equiv Y_{t-j}$ for all j, the model is given by

$$Y_t = m(Y_{t-1}, Y_{t-2}, \ldots) + \varepsilon_t \tag{2.25}$$

and whether the stationary solution $\{Y_t\}$ indeed exists is an important question. In the study of a class of general nonlinear AR(d) models, Duflo (1997) and Götze and Hipp

(1994) assumed what is called the Lipschitz mixing condition (or the strong contraction condition), which is essentially (2.24) replaced by finite d-sum on the right hand side. In our context, Assumption B7 plays a similar role; Doukhan and Wintenberger (2008) showed that (2.24) with $\sum_{j=1}^{\infty} c_j < 1$, is sufficient for the existence of a stationary solution: for some measurable f,

$$Y_t = f(\varepsilon_t, \varepsilon_{t-1}, \ldots), \tag{2.26}$$

where ε_t is an i.i.d. sequence. Wu (2011) arrived at the same conclusion under the assumption of $\sum_{j=1}^{\infty} c_j = 1$; the specific restrictions on c_j are chosen to reflect their findings, despite the fact that we are not restricting the error process $\{\varepsilon_t\}$ to be an independent sequence in our model setup.

Before we proceed, we remark that from now on the rate condition stipulated in (2.19) is slightly strengthened as follows (and Assumption A2 is modified accordingly):

$$v_2(r)^{1/2}[\varphi_x(\underline{h}\lambda)]^{-1}n^{1/2} \to 0 \qquad as \;\; n \to \infty. \tag{2.27}$$

(2) The 'variance component'

We now move on to the second chunk of assumptions that are concerned with the 'variance part'. As before, vectors Z and z are taken to mean $(\phi_1^{-1}X_1, \phi_2^{-1}X_2, \ldots)^\intercal$ and $(\phi_1^{-1}x_1, \phi_2^{-1}x_2, \ldots)^\intercal$, respectively, where the vector $x = (x_1, x_2, \ldots)^\intercal$ is the point at which estimation is made, and ϕ_j's are the coefficients in Assumption B8.

Assumptions D

D1. *The induced probability measure P_{z-Z} is dominated by the measure P_Z, and its Radon-Nikodym density $dP_{z-Z}/dP_Z =: p^*$ is continuous and is bounded away from zero at $0 \in \mathbb{R}^\infty$; i.e., $p^*(0) > 0$.*

D2. *The distribution F of X_s^2, where each X_s is the marginal regressor, is regularly varying near zero with strictly positive index $(-\rho) > 0$.*

D3. *Further to B8, the bandwidth satisfies $h_j = j^p h$ (i.e. $\phi_j = j^p$) with $p \in \Pi(c, \beta)$, where*

$$\Pi(c, \beta) = \left\{ p : \sum_{j=1}^{\infty} c_j j^{p\beta} < \infty, \frac{1}{2} < p \right\}.$$

D4. *The conditional variance* $\mathrm{var}[Y_t|X_t = u] = \sigma^2(u)$ *is continuous in some neighbourhood of x; i.e.* $\sup_{u \in \mathcal{E}(x,\underline{h}\lambda)}[\sigma^2(u) - \sigma^2(x)] = o(1)$. *Similarly, the cross-conditional moment $E[(Y_t - m(x))(Y_s - m(x)|X_t = u, X_s = v] = \sigma(u, v)$, $t \neq s$ is continuous in some neighbourhood of (x, x).*

D5. $R_{nt} := (EK_1)^{-1}\{K_t(Y_t - m(x)) - EK_t(Y_t - m(x))\}$ *belongs to the domain of attraction of a normal distribution.*

Remark. Assumption D1 is concerned with a transition of the shifted small ball probability to the centred small deviation (whose asymptotic behaviour is more accessible), see Mas (2012). The explicit form of the derivative (and hence of the relationship between the two probabilities) cannot be easily computed in general. Nonetheless, in the special case of the Gaussian process Z with covariance operator Σ it is known by Sytaya (1974) and Zolotarev (1986) that

$$P\Big(\|z - Z\| \leq \epsilon\Big) \simeq P\Big(\|Z\| \leq \epsilon\Big) \exp\Big\{ -\frac{1}{2}\|\Sigma^{-1/2}z\|^2\Big\} \quad \text{as} \quad \epsilon \to 0. \tag{2.28}$$

The reader is directed to Li and Shao (2001) for detailed discussion on this asymptotic equivalence relation. Note that Σ can be expressed in terms of the a_j constants (in Assumption C), which govern the dependence between the marginal regressors and the bandwidth weights ϕ_j:

$$\mathrm{cov}(Z) = \Sigma = (DA)(A^*D), \tag{2.29}$$

where $A = (a_{ij}) = (a_{i-j})$ and $D = \mathrm{diag}(\phi_1, \phi_2, \ldots)$.

Condition D2 is equivalent to saying that

$$\lim_{x\to\infty} \frac{F(1/(\gamma x))}{F(1/x)} = \gamma^\rho,$$

where ρ is the index of variation which is strictly negative. Under this condition, Dunker, Lifshits and Linde (1998, cf. Conditions I and L) derived the explicit behaviour of the small ball probability. We require the function $F(1/x)$ to be regularly varying in order to ensure that the small ball probability is *well-behaved* near infinity in the asymptotic sense. Since only those functions having strictly negative ρ satisfy the condition, the distribution F of the squared regressor must be such that $F(1/x)$ decreases (as $x \to \infty$) at a *reasonable speed*. By reasonable we mean that the relative weight of decrease follows a power law, and the variation should be continuous. A large class of common distributions satisfies this condition; for example: the Gamma, Beta, Pareto, Exponential, Weibull, and also the Chi-squared distribution (in which case each X_s is Gaussian). Indeed, both D1 and D2 hold under Gaussianity (e.g. when condition C2 is assumed).

The specific bandwidth increment condition assumed in D3 is one framework under which the explicit behaviour of the small ball probability can be specified (cf. Dunker et al. (1998)). In the exceptional case of static regression where the regressors form an i.i.d. sequence, the probability can also be derived when the weights are of an exponential type (i.e. $h_j = e^j h$) up to an unknown function, or are non-increasing in a particular manner (cf. Gao et al. (2003)) similar to the polynomial decay. In this chapter however, we shall confine our attention to the case of the polynomial law for expositional simplicity and consistency of presentation, since the asymptotic behaviour of the small ball is not yet known in the dependent case for choices other than the polynomial decay as in D3. The standard conditions in D4 are assumed to deal with the asymptotics of the variance and covariance terms. The last condition is imposed to establish the self-normalized CLT without assuming higher moment conditions; relevant discussions can be found for example in de la Peña et al. (2009). The condition is not affected by the dependence structure of R_{nt} as the property is inherited to the approximated sum in the Bernstein's blocking procedure; see (2.78) for details.

In practice, we would require some ordering for the marginal regressors in the static regressions case A1, since the influence of marginals is set to decrease via the bandwidth adjustments as discussed just above. One practical way of doing this is to rank them in the order of goodness of fit, or the contribution that each marginal regressor makes in the estimation. For example, one could evaluate the sample correlations between Y_t and $\widehat{\mathrm{E}}(Y_t|X_{jt})$, where X_{jt} is a marginal covariate and $\widehat{\mathrm{E}}(Y_t|X_{jt})$ is a kernel estimate of the univariate marginal regression, and order them according to the computed correlations. This way one can line up the marginal regressors in the order of their relative importance. This method is motivated by the Kernel Sure Independence Screening (KSIS) approach in Chen, Li, Linton and Lu (2017), and the reader is referred to their paper for further details.

We are now able to derive the following results for the bias and variance components using the decomposition (2.21), Assumptions B7, B8 and C, and Corollary 2.1:

$$\mathcal{B}_n(x) := \left[E\widehat{m}_2(x) - m(x)\right] \leq h^\beta \lambda^\beta \sum_{j=1}^{\infty} c_j j^{p\beta} \tag{2.30}$$

$$\mathcal{V}_n(x) := \mathrm{var}\left[\widehat{m}_2(x)\right] \simeq \frac{\sigma^2(x)\xi_2}{n\varphi_x(\underline{h}\lambda)\xi_1^2}, \tag{2.31}$$

where λ and $\widehat{m}_2(\cdot)$ are as in (2.5) and (2.20), respectively. Formal derivation is done in 2.7.2 of the appendix. We next present the CLT of our estimator.

Limiting distribution under independence of regressors (A1 & C1)

We first consider the situation in which there is a set of independent exogenous regressors in the static regression context. That is, when marginal regressors X_s are independent to each other and are identically distributed (i.e. satisfies Assumption C1). It is then natural to assume that the samples follow Assumption A1, since having A2 when C1 holds means $Y_t, Y_{t-1}, \ldots$ are independent.

$X_j^2 \sim F$ i.i.d.	ρ	$\lim_{x\to\infty} \ell(x) = C_\ell^{-2}$	ζ
Uniform(1,b)	-1	1	n/a
Gamma(α, β)	$-\alpha$	$\beta^\alpha \alpha^{-1} \Gamma(\alpha)^{-1}$	$\frac{\alpha\pi\beta^{-1/2p}}{\sin(\pi/2p)}$
Exponential(η)	-1	η	$\frac{\pi\eta^{-1/2p}}{\sin(\pi/2p)}$
Weibull(α, β)	$-\alpha$	β	n/a
Pareto(θ, μ)	-1	μ/θ	n/a
χ_1^2	$-1/2$	$(2/\pi)^{1/2}$	$\frac{\pi 2^{(1-2p)/2p}}{\sin(\pi/2p)}$

Table 2.1: Examples of the key constants for some common distributions

In this case, the asymptotic normality can be established for regressors that follow a wide range of different distributions. Recall that under Assumption D2, the distribution function F (of X^2) is regularly varying with the index of variation $\rho < 0$. Then, by the characterization theorem of Karamata (1933) (see e.g. Feller (1971)), there always exists a slowly varying function $\ell(x)$ satisfying

$$F(1/x) = x^\rho \ell(x). \tag{2.32}$$

Now fix some p, the order of increment constant for bandwidth in Assumption D3, and denote by $\mathcal{L}(t)$ the Laplace transform of X^2. We then define the following constants:

$$C_\ell = \lim_{\delta\to 0}\left[\ell^{-1/2}\left(\delta^{-\frac{4p}{2p-1}}\right)\right], \quad \zeta = -\int_0^\infty \frac{u^{-1/2p}\mathcal{L}'(u)}{\mathcal{L}(u)}du$$

$$C^* = \frac{(2\pi)^{(1+2p\rho)}(2p-1)}{\Gamma^{-1}(1-\rho)\cdot(2p)^{\frac{2p(\rho+2)-1}{2p-1}}}\cdot\zeta^{\frac{2p(1+\rho)}{2p-1}}, \qquad C^{**} = (2p-1)\cdot\left(\frac{\zeta}{2p}\right)^{2p/(2p-1)}$$

$$\kappa_0(K,p,F) = C^{**}\lambda^{-\frac{2}{2p-1}} \quad \text{and} \quad \kappa_1(K,p,F) = \frac{C^* C_\ell \xi_2}{p^*(0)\xi_1^2\lambda^{\frac{1+2\rho p}{2p-1}}},$$

where $\Gamma(\cdot)$ is the Gamma function, ξ_1 and ξ_2 are the constants specified in (2.12) (which simplify in case of uniform kernel for example), λ is the upper bound of the support of the kernel, and $p^*(\cdot)$ is the Radon-Nikodym derivative in D1. The underlying arguments for the formulation of these constants can be found in Dunker, Lifshits and Linde (1998). To aid the exposition, we compute and present the constants for some common, regularly varying distributions in Table 2.1.

The main result of this subsection now follows. The theorem gives the limiting distribution of the estimator (2.18) under cross sectional independence with respect to mixing sample data.

Theorem 2.2. *Suppose that B2-B8 and D1-D4 hold. Let the marginal regressors X_s*

satisfy Assumption C1. Then the estimator (2.18) based on the sample observations $\{Y_t, X_t\}_{t=1}^n$ satisfying A1 is asymptotically normal with the following limiting distribution:

$$\sqrt{nh^{\frac{1+2\rho p}{2p-1}} \exp\left(-\kappa_0 h^{-\frac{2}{2p-1}}\right)} \left(\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\right) \Longrightarrow N\left(0, \kappa_1 \sigma^2(x)\right), \tag{2.33}$$

where $\mathcal{B}_n(x) = O(h^\beta)$ is the 'bias part' in (2.30) and $\sigma^2(x)$ is the conditional variance defined in D4.

Limiting distribution under Gaussianity & dependence of regressors (A & C2)

The independence condition between the regressors assumed in the previous section can be relaxed to allow some mild dependence specified in Assumption C2. In doing so, we make use of the result derived in Hong, Lifshits and Nazarov (2016, Theorem 1.1), where the asymptotics of the small deviation probability of Gaussian dependent sequences was investigated. This setting not only grants sufficient flexibility in the static regression case (Assumption A1), but moreover allows one to compute the distributional result for the dynamic regression context, where the regressor vector consists of time lags of the response or a covariate with dependence structure as described in Assumption A2. The price we have to pay for this flexibility is the Gaussianity restriction on the regressors.

With reference to Table 2.1 above, we can easily compute the constants C^* and C^{**} for the Gaussian case, denoted C^*_G and C^{**}_G respectively, as follows:

$$C^*_G = \frac{(2\pi)^{(1-p)}(2p-1)}{2 \cdot (2p)^{\frac{3p-1}{2p-1}}} \cdot \left[\frac{\pi 2^{(1-2p)/2p}}{\sin(\pi/2p)}\right]^{\frac{-p}{2p-1}}, \qquad C^{**}_G = \frac{2p-1}{2}\left(\frac{\pi}{2p\sin\frac{\pi}{2p}}\right)^{\frac{2p}{2p-1}}.$$

For the square summable sequence a_j in (2.23) define

$$C_{\mathcal{A}} = \left[\frac{1}{2\pi}\int_0^{2\pi}\left|\sum_{j=0}^{\infty} a_j \exp(\mathrm{i}js)\right|^{1/p} ds\right]^p \quad \text{and} \quad \kappa_2(K,p,a) = \frac{C^*_G C_\ell \xi_2 \xi_1^{-2}}{e^{-\frac{1}{2}\|\Sigma^{-1/2}z\|_2^2}(C_{\mathcal{A}}\lambda^{-1})^{\frac{p-1}{2p-1}}}$$

where $z = (z_j) = (j^{-p}x_j) = D^{-1}x$. Recall that for the uniform (Box) kernel $\xi_2 = \xi_1^2$, so they cancel out in κ_2. It is worth noting that $C_{\mathcal{A}}$ is a function of the spectral density of the MA(∞) process $\{X_{jt}\}_j$. Let $\kappa_0' = C^{**}_G C_{\mathcal{A}}^{2/(2p-1)}\lambda^{-2/(2p-1)}$.

With other constants defined as before, we now have the following asymptotic normality for the case of dependent regressors. We reiterate that the result covers both the static and dynamic regressions context (A1 and A2). Comparing the Gaussian cases under C2 and C1 (cf. Theorem 2.2), we notice that it is only the constant factor $C_{\mathcal{A}}$ that is additionally involved. A consequence of this difference in terms of the rate of convergence is discussed in the subsection that follows.

Theorem 2.3. *Suppose Assumptions B2-B8, D3, and D4 hold. Let the regressor $X = (X_1, X_2, \ldots)^\intercal$ be jointly normally distributed with zero mean and the covariance operator Σ, and satisfies C2. Then, the estimator (2.18) with respect to sample $\{Y_t, X_t^\intercal\}_{t=1}^n$ satisfying either A1 or A2 is asymptotically normal with the following limiting distribution:*

$$\sqrt{nh^{\frac{1-p}{2p-1}} \exp\left(-\kappa_0' h^{-\frac{2}{2p-1}}\right)} \left(\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\right) \Longrightarrow N\left(0, \kappa_2 \sigma^2(x)\right), \tag{2.34}$$

where $\mathcal{B}_n(x) = O(h^\beta)$ is the 'bias part' in (2.30) and $\sigma^2(x)$ is the conditional variance defined in D4.

Remark. The additional constant $C_{\mathcal{A}}$ is a function of the sequence a_j, and represents the dependence structure allowed between the regressors. This suggests an interesting finding that says allowing for dependence does not incur much penalty; we conjecture that similar conclusion would hold for regressors of different distributions than Gaussian, but leave it for future studies. The exponential term in the denominator of the asymptotic variance arises from the asymptotic equivalence relationship between the shifted and non-shifted small deviation for ℓ_2-valued Gaussian variables, cf. (2.28).

In both frameworks of independent regressors and dependent Gaussian regressors we are able to construct the self-normalised central limit theorem; define

$$\Delta_n^2(x) = \sum_{t=1}^n \left(\sum_{s=1}^n K_s\right)^{-2} \left[K_t\big(Y_t - \widehat{m}(x)\big)\right]^2, \tag{2.35}$$

where $K_t := K(\|H^{-1}(x - X_t)\|)$ as before.

Corollary 2.2. *Further to the conditions assumed either in Theorem 2.2 or Theorem 2.3, suppose that Assumption D5 holds. Then the following central limit theorem holds*

$$\Delta_n^{-1}(x)\Big(\widehat{m}(x) - m(x) - \mathcal{B}_n(x)\Big) \Longrightarrow N(0, 1),$$

where $\Delta_n(x)$ is the square root of (2.35).

This self-normalized limit distribution gives (pointwise) confidence intervals for $\widehat{m}(x)$, which can be used as a basis for conducting standard statistical inference.

2.3.3 Optimal Bandwidth

We now discuss the issue of bandwidth optimality. As in the finite-dimensional framework, there is a bias-variance trade-off. As the bandwidth goes up, the variance gets smaller while the bias increases, and vice versa. Therefore we search for the optimal bandwidth h_{opt} that balances the order of those two quantities.

We first suppose that $p \in \Pi(c, \beta)$, cf. D3, is given. In the i.i.d. case with Gaussian regressor we have

$$h^\beta \sim \sqrt{\frac{\exp\left(\kappa_0 h^{-2/(2p-1)}\right)}{n h^{\frac{1-p}{2p-1}}}}, \tag{2.36}$$

so that

$$\left[2\beta + \frac{1-p}{2p-1}\right] \cdot \log h - \kappa_0 h^{-\frac{2}{2p-1}} \sim -\log n.$$

Taking $h \sim (\log n)^a$ for some $a < 0$ balances the leading terms on both sides:

$$\left[2\beta + \frac{1-p}{2p-1}\right] \cdot a \cdot \log\log n - \kappa_0 (\log n)^{-\frac{2}{2p-1} \cdot a} \sim -\log n. \tag{2.37}$$

The explicit order a that solves (2.37) can be expressed in terms of n, β and p. Writing $\vartheta := [2\beta + (1-p)/(2p-1)]$ and $\chi := 2/(2p-1)$ for notational simplicity, and solving for a we have

$$a_{opt} = \frac{\vartheta \cdot \mathcal{W}\left(\frac{\chi}{\vartheta} \cdot \kappa_0 \cdot n^{\chi/\vartheta}\right) - \chi \log n}{\vartheta\chi \cdot \log\log n}, \tag{2.38}$$

where $\mathcal{W}(y)$ is the Lambert W function (see e.g. Olver et al. (2010)), which returns the solution x of $y = x \cdot e^x$. From (2.38) the optimal bandwidth $h_{opt} \sim (\log n)^{a_{opt}}$ follows, in which case the asymptotic root mean squared error is of the order $(\log n)^{\beta a_{opt}}$.

Remark. We can look for the optimal bandwidth for the cases of non-Gaussian regressors by following exactly the same procedure as above; tedious details are omitted here. In the dependent case (i.e. under C2), apart from having κ_0' instead of κ_0 everything remains the same, and this does not affect the *asymptotics* as $C_{\mathcal{A}}$ is a fixed quantity. Regarding the solution in (2.38), since the mapping $x \mapsto x \cdot e^x$ is not an injection, the solution may be multi-valued on the negative domain, i.e. $y < 0$. This does not happen in (2.38) provided $\beta \geq 1/4$ (however big p is), because $(1-p)/(2p-1)$ is bounded away from $-1/2$; in this case, the coefficient of the double logarithmic term in (2.37) is strictly smaller or equal to zero.

Since the log terms dominate the double logarithm in (2.37) as the sample size n increases, it can be readily expected that the optimal value of a in (2.38) converges to a limit in such a way that the leading orders are balanced. Below we introduce without formal justification a trivial result that gives the lower bound (infimum) of the optimal bandwidth (and hence of the optimal rate that balances the bias and variance). We remark that the result below holds for other choices of the distribution of the regressors, and also for the case of dependent regressors as in Assumption C2, since the exponent of the leading term $-2/(2p-1)$ remains invariant as it was shown in (2.33) and (2.34). Nonetheless, it is worth noting that although the order of the convergence rate remains the same, the difference in associated constants (e.g. κ_0 and κ_0') does make the speed at which a_{opt} converge to the limit in (2.39).

Corollary 2.3. *For any fixed choice of $p \in \Pi(c,\beta)$ and the distribution F of X^2 satisfying Assumption D2, the order of the optimal bandwidth a_{opt} satisfies*

$$a_{opt} \downarrow \left(-\frac{2p-1}{2}\right) \quad as \quad n \to \infty, \tag{2.39}$$

which suggests that the lower bound of the optimal bandwidth is given by

$$(\log n)^{-\frac{2p-1}{2}} \preceq h_{opt} \sim (\log n)^{a_{opt}}. \tag{2.40}$$

This result tells us the best possible performance we can expect from the optimal bandwidth. Because $n^k(\log n)^{-(2p-1)/2} \to \infty$ for any real $k > 0$, it follows that we cannot possibly estimate the regression function at a polynomial rate. This result is consistent with and complements the findings of Masry (2012, Theorem 3), which were obtained under the assumption of independence of regressors. The paper also considered the case where the bandwidth grows exponentially: $\phi_j \succeq \exp(j^q)$ for some $q > 0$. Then for $h_j = \phi_j h$, his result suggests $\exp[-(\log n)^{2q/(2q+1)}] \preceq h_{opt} \sim \exp[a_{opt} \cdot (\log n)^{b_{opt}}]$. Therefore, the performance is better in general in this case, although obviously a polynomial rate of convergence still cannot be attained. It is not clear what will happen when the regressors are allowed to be dependent in the sense of Assumption C2, since the behaviour of the small ball probability for dependent sequence is not known for the case of exponentially decaying weights.

Returning back to Corollary 2.3, we emphasize that the arguments are true for any $p \in \Pi(c,\beta)$. Let $p_{\max} = \sup_{p\in\Pi(c,\beta)} p$. Then a lower bound on the optimal bandwidth (over all p) is $(\log n)^{-\left(p_{\max}-\frac{1}{2}\right)}$. For example, when $c_j = (1/2)j^{-2}$ we have $p_{\max} = 1/\beta$. Unfortunately, it is generally the case that $p_{\max} \notin \Pi(c,\beta)$, in which case the lower bound is not quite achievable by our method.

Remark. Regarding bandwidth selection, one possibility is the Bayesian bandwidth selection methods like proposed in Zhang, King, and Hyndman (2006). We take as prior for h the density proportional to $1/(1+\lambda h^2)$ and as prior for $p - 1/2$ the density of a $\chi^2(w)$ random variable. The hyperparameters λ, w may be chosen by experimentation. The priors are combined with a Gaussian (least squares) density to deliver a posterior for the bandwidth. The reader is referred to Section 2.5 for further discussions on the issue of bandwidth choice.

2.3.4 Uniform consistency

Uniform consistency of the Nadaraya-Watson estimator was first studied by Nadaraya (1964, 1970) and subsequently by numerous others. To mention few early papers, Devroye (1978) weakened the regularity conditions required in the previous papers, and Robinson (1983) proved uniform consistency for dependent sample data. In the functional statistics literature, uniform consistency of kernel estimators for conditional mean is established only with respect to i.i.d. sample data so far (see for example Ferraty et al. (2010), Ferraty et al. (2011), Kudraszow and Vieu (2013), and Kara-Zaïtri et al. (2017)) as far as the authors are concerned.

In this section, we show uniform consistency of our estimator under the (suitably modified) regularity conditions assumed in the previous sections. We start by introducing the notion of Kolmogorov's entropy below. For some of its ealier discussions in statistics literature, the reader is referred to Yaracos (1985) and Mammen (1991).

DEFINITION 2.5. *Given some $\eta > 0$, let $L(S, \eta)$ be the smallest number of open balls in E of radius η needed to cover the set $S \subset E$. Then Kolmogorov's η-entropy is defined as* $\log L(S, \eta)$.

This quantity will be used in explaining the topological restrictions we adopt to suitably accommodate infinite dimensionality. The definition implies the dependence of Kolmogorov's entropy both on the nature of the space under study and the measure of proximity. It will be shown later in this section that the entropy is closely related to the rate of convergence of the estimator, in particular, to the penalty incurred on the rate in the uniform case. It is well known that the regression function cannot be estimated uniformly over the entire space, e.g. Bosq (1996). In our infinite dimensional framework, the infinite sequence spaces, if unrestricted, cannot be covered by a finite number of balls, and that $L(S, \eta) = \infty$. We propose to consider uniform consistency over a subset of $\mathbb{R}^\infty$, whose effective dimension is truncated and is increasing in sample size n. In particular, we define the set

$$S_\tau := \left\{ u | (u_i)_{i \in \mathbb{Z}^+}, u_j = 0 \text{ for all } j > \tau, \|u\|_\infty \leq \lambda \right\} \subset \mathbb{R}^\infty, \tag{2.41}$$

where $\tau = \tau_n$ is some increasing sequence and λ is fixed, and consider uniform consistency over this compact set. Then Kolmogorov's entropy of the set S_τ is given as follows:

Lemma 2.2. *Kolmogorov's η-entropy of S_τ defined in (2.41) with $\tau = \tau_n (\to \infty)$ and $\lambda > 0$ is*

$$\log L\big(S_\tau, \eta\big) = \log \left[\left(\frac{2\lambda\sqrt{\tau}}{\eta} + 1 \right)^\tau \right]. \tag{2.42}$$

Remark. (2.42) is in line with common intuition; as the effective dimension τ increases, the number of balls (with some fixed radius) required to cover the set tends to infinity. Lemma 2.2 can be shown by exploiting the splitting technique and then by covering the polyhedron of increasing dimension. See appendix for details. Note that for some fixed λ and $\eta = \eta_n$, Kolmogorov's entropy $\log L(S_\tau, \eta)$ is of order $(\tau \log \tau - \tau \log \eta)$.

Considering the definition of the set S_τ, in the sequel (with a slight abuse of notation) we take X to denote the regressor, but with zeros after its $\tau^{th}(= \tau_n \to \infty$ as $n \to \infty)$ entry; that is, $X = (X_1, X_2, \ldots, X_\tau, 0, 0, \ldots)^\intercal$ (so that the original X is recovered as $n \to \infty$). Also, the regression operator and the estimator with respect to this truncated regressor are denoted by $m_\tau(\cdot)$ and $\widehat{m}_\tau(\cdot)$, respectively. All assumptions, including the additional one to follow below, are understood to hold under these modifications.

Assumption E

E. *For sufficiently large n, Kolmogorov's η-entropy $\log L(S_\tau, \eta)$ satisfies*

$$\frac{(\log n)^{8+2\epsilon}}{n\varphi_x(\underline{h})} \le \log L(S_\tau, \eta) \le \frac{\sqrt{n\varphi_x(\underline{h})}}{(\log n)^{1+\epsilon}} \quad \text{for some} \;\; \epsilon \in (0, 1/2). \tag{2.43}$$

Furthermore, $0 < \varphi_x(\underline{h}) \preceq h < \infty$ *and* $(\log n)^2/(n\varphi_x(\underline{h})) \longrightarrow 0$ *as* $n \to \infty$.

Remark. The first part of Assumption E specifies the rate at which Kolmogorov's entropy should behave with sample size n (hence in dimension $\tau = \tau_n$). From the upper and lower bound it readily follows that $n\varphi(h)$ must be of order larger than $(\log n)^{6+2\epsilon}$. This assumption is sufficiently general. For example, in view of the bias-variance optimal bandwidth suppose $h \sim (\log n)^{-(2p-1)/2}$ so that $n\varphi(h) \sim (\log n)^{(2p-1)\beta}$. In this case, assumption (2.43) is valid as long as p is moderately large enough relative to $\beta \le 1$ in such a way that $6 + 2\epsilon \le (2p-1)\beta$. The second part is standard; in particular, the last condition straightforwardly follows by (2.43) and only slightly strengthens the bandwidth condition in Assumption B2.

For uniform consistency we impose a stronger condition on mixing coefficients. From hereafter, by A1$'$ and A2$'$ we mean Assumptions A1 and A2 but with the arithmetic mixing rate condition strengthened to the following exponential mixing condition (cf. Definition 2.1):

$$\alpha(r) \le \exp(-\varsigma r^{\gamma_2}) \tag{2.44}$$

where $\varsigma > 1$ and γ_2 is a positive constant such that $\gamma := 1/(\gamma_1^{-1} + \gamma_2^{-1}) \ge 1$, with γ_1 defined as in Assumption B4$'$. When the response is assumed to be bounded (i.e.

$|Y_t| \le C$), γ_1 may be taken to be ∞ so that $\gamma_2 = \gamma \ge 1$. This stronger mixing condition enables us to obtain exponential bounds that decay fast enough, thereby accommodating uniformity, see appendix for details. We hope to expect the same conclusion in this section to hold under the arithmetic mixing condition we previously assumed, once some suitably sharper exponential inequality becomes available. In line with the modification on the mixing rate above, we also impose a slightly stricter condition on the response:

B4′. *The response Y_t is satisfies the following tail condition: There exists some positive constant γ_1 and C such that $P(|Y_t| > u) \le C \exp(1 - u^{\gamma_1})$ for any $u > 0$.*

For example, a Gaussian random variable satisfies B4′ with $\gamma_1 < 2$. The condition is also satisfied by many unbounded variables and all those bounded ones. The main result of this section now follows.

Theorem 2.4. *Suppose that Assumptions B2, B3, B4′, B5-B8, D1-D3 and E hold. Let the marginal regressors X_s satisfy C1, and take $\tau = \tau_n \sim (\log n)$. Then the estimator $\widehat{m}_\tau(\cdot)$ with respect to sample observations $\{Y_t, X_t\}_{t=1}^n$ satisfying A1′ is uniformly consistent for $m(x) = m(x_1, x_2, \ldots)$ over S_τ:*

$$\sup_{x \in S_\tau} \left| \widehat{m}_\tau(x) - m_\tau(x) \right| = O_P\left(h^\beta + \sqrt{\frac{(\log n)^2 \exp\left(\kappa_0 h^{-2/(2p-1)}\right)}{n h^{\frac{1+2p\rho}{2p-1}}}} \right). \tag{2.45}$$

If alternatively X_s is Gaussian and satisfies C2, then the same conclusion holds with respect to sample observations satisfying either A1′ or A2′, in which case κ_0 is replaced by κ_0' in Theorem 2.3.

Remark. We may choose the optimal bandwidth as before; following the same arguments in the pointwise case, choosing $h \sim (\log n)^a$ and solving for n gives

$$a_{opt} = \frac{\vartheta \cdot \mathcal{W}\left[\frac{\chi}{\vartheta} c \exp(-\frac{\chi}{\vartheta} 2 \log\log n + \chi \log n)\right] + 2\chi \log\log n - \chi \log n}{\vartheta \chi \log\log n}. \tag{2.46}$$

And because the order of the leading terms is $(\log n)^{-(2p-1)/2}$ as in the pointwise case, it is straightforward to see that the lower bound of the optimal bandwidth in Corollary 2.3 still continues to hold; that is, $h_{opt} \succeq (\log n)^{-(2p-1)/2}$. This is again invariant to the choice of distribution F of the squared regressor. It is important to note that as before, potential cross-sectional dependence between marginal regressors and also their distributional properties are represented via c, the collection of constants that appear inside the exponential terms either in (2.33) and (2.34).

The results altogether give the optimal rate of convergence of our estimator as follows.

Corollary 2.4. *Suppose conditions assumed in Theorem 2.4 hold. Upon choosing $h \sim (\log n)^{a_{opt}}$, where a_{opt} is as defined in (2.46), we have*

$$\sup_{x \in S_\tau} \left| \widehat{m}_\tau(x) - m_\tau(x) \right| = O_P\Big([\log n]^{\beta \cdot a_{opt}} \Big). \tag{2.47}$$

In the pointwise case the same result in Corollary 2.4 trivially holds, but with the different optimal a_{opt}; it is as given in (2.38). In that case this rate of convergence is minimax optimal in view of Theorem 3 of Mas (2012). Although both a_{opt} converge to $-(2p-1)/2$, the speed at which they converge is different as can be seen in the example in Figure 2.1 below.

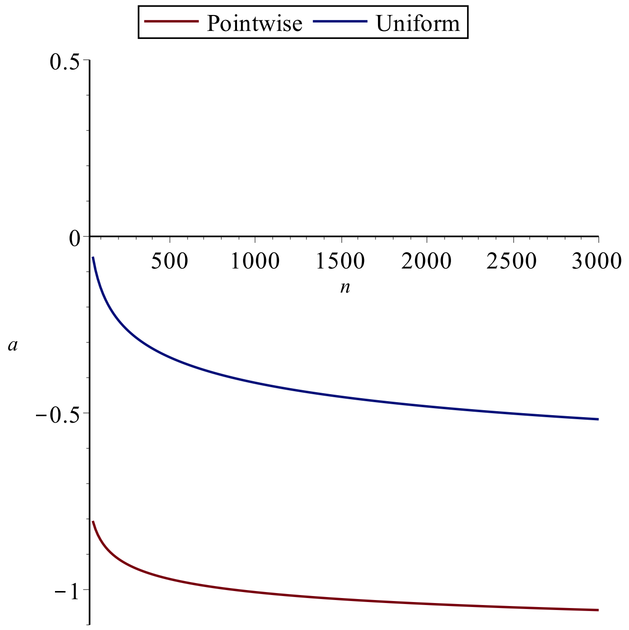


Figure 2.1: $a_{opt} = a_{opt}(n)$ for $\beta = 1$ and $p = 2$

2.4 Application to the Risk Return Relationship

The relation between the expected excess return on the aggregate stock market - the so called “equity risk premium” - and its conditional variance has long been the subject of both theoretical and empirical research in financial economics. The risk-return relation is an important ingredient in optimal portfolio choice, and is central to the development of theoretical asset-pricing models aimed at explaining a host of observed stock market patterns. Asset pricing models generally predict a positive relationship between the risk premium on the market portfolio and the variance of its return. In an influential paper, Merton (1973) obtained very simple restrictions albeit under somewhat drastic assumptions; he showed in the context of a continuous time partial equilibrium model that

$$\mu_t = E[(r_{mt} - r_{ft})|\mathcal{F}_{t-1}] = \gamma \text{var}[(r_{mt} - r_{ft})|\mathcal{F}_{t-1}] = \gamma\sigma_t^2, \tag{2.48}$$

where r_{mt}, r_{ft} are the returns on the market portfolio and risk-free asset respectively, while $\mathcal{F}_{t-1}$ is the market wide information available at time $t-1$. The positive constant γ is the Arrow–Pratt measure of relative risk aversion. The linear functional form actually only holds when σ_t^2 is constant; otherwise μ_t and σ_t^2 can be nonlinearly related, Gennotte and Marsh (1993). Further examples with a positive risk return trade-off include the external habit model of Campbell and Cochrane (1999) and the Long Run Risks model of Bansal and Yaron (2004). However, a negative risk-return relation is not inconsistent with (a general enough) equilibrium, Backus and Gregory (1993). Unfortunately, the empirical evidence on the risk-return relation is mixed and inconclusive. Ghysels, Santa-Clara, and Valkanov (2005), Bali and Peng (2006), Lundblad (2005), Pástor, Sinha, and Swaminathan (2008), and Ludvigson and Ng (2007) find a positive risk-return relation, while Campbell (1987), Glosten, Jagannathan, and Runkle (1993), Harvey (2001), and Lettau and Ludvigson (2003) find a negative relation. Still others find mixed and inconclusive evidence like French, Schwert, and Stambaugh (1987), Nelson (1991), Campbell and Hentschel (1992), Linton and Perron (2003), and Whitelaw (1994). Scruggs (1998) and Guo and Whitelaw (2006) document a positive trade-off within specifications that facilitate hedging demands. However, Scruggs and Glabadanidis (2003) find that this partial relationship is not robust across alternative volatility specifications.

As already mentioned in the beginning of this chapter, the main difficulty in estimating the risk-return relation is that neither the conditional expected return nor the conditional variance of the market is directly observable. The contradictory findings of the above studies are mostly the result of differences in the specifications and approaches to modeling the conditional mean and variance. Pagan and Ullah (1988), and Pagan and Hong (1990) initiated the use of nonparametric methods in this setting. The latter paper argued that the risk premium μ_t and the conditional variance σ_t^2 are highly nonlinear functions of the past whose form is not captured by standard parametric GARCH–M models. They estimated $E(r_{mt}-r_{ft}|r_{m,t-1},\ldots,r_{m,t-p})$ and $\text{var}(r_{mt}-r_{ft}|r_{m,t-1},\ldots,r_{m,t-p})$ nonparametrically, where $p \in \{1,4\}$, finding evidence of considerable nonlinearity. They then estimated δ from the regression

$$r_{mt} - r_{ft} = \delta\sigma_t^2 + \eta_t, \tag{2.49}$$

by OLS and IV methods, finding a negative but insignificant δ. There are a number of drawbacks with the Pagan and Hong (1990) approach. Firstly, as aforementioned in the introduction, the conditional moments are calculated using a finite conditioning set. This greatly restricts the dynamics for the variance process. Secondly, they only test for linearity of the relationship between μ_t and σ_t^2; this seems to be somewhat restrictive in view of earlier findings. Linton and Perron (2003) considered the model where σ_t^2 was a parametrically specified CH process (with dependence on the infinite past) but $\mu_t = \varphi(\sigma_t^2)$ for some function φ of unknown functional form. They proposed an estimation algorithm but did not establish any statistical properties. They found some evidence of a nonlinear relationship. Conrad and Mammen (2008) develop the theory of estimation and inference for this model. Christensen, Dahl, and Iglesias (2012) developed the theoretical framework by considering volatility models that are driven by observable shocks so that a full theory can be given. Escanciano, Pardo-Fernández and Van Keilegom (2017) consider a more general class of semiparametric models. Under the semi-strong form of the efficient market hypothesis prices contain all relevant information and so the risk premium and risk themselves can be expressed in terms of only the past history of prices. We shall use this assumption to obviate the omitted variables/endogeneity issues that have limited previous applications in this area.

2.4.1 Empirical study on the US stock market

We apply our methods to the daily risk premium on the value weighted S&P500 index — the total return on the index minus the returns on T-bills[6], denoted Y_t — over the period 04 January 1950 to 30 August 2017, a total of 17,025 observations. The whole time period is divided into 5 subperiods: 1950:01:04-1963:02:21, 1963:02:25-1976:05:04, 1976:05:05-1989:05:24, 1989:05:25-2002:06:24, and 2002:06:25-2017:08:30, to see if there is any variation in the ex-post risk and return by decades. Except for the last subperiod where there are 3824 observations, the other five each contains 3300 observations. We suppose that both the conditional mean and variance of Y_t, denoted μ_t and σ_t^2, are unrestricted nonparametric functions of the entire information set. We estimate them for $p = 4$ and 12 at the points $X_t = (Y_{t-1}, Y_{t-2}, \ldots, Y_1, 0, 0, \ldots)$. The uniform kernel $K(\|u\|) = 1_{[0,1]}(\|u\|)$ is used, and the bandwidth sequence h of 0.00035 and 0.000125 are used for $p = 4$ and $p = 12$, respectively. These bandwidths are in accordance with the selection methods we propose below in the end of this section.

Table 2.2 reports some summary statistics of the nonparametric estimates $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ for $p = 4$ over the full period (1950-2017). We present the mean, standard deviation, skewness, kurtosis and the fitted AR(1) coefficients. The estimated conditional variance shows high persistence. Table 2.2 may be compared with Table I of Bali and Peng (2006), where they report similar descriptive statistics for their realized, GARCH, and implied volatility estimates computed using 5-minute high frequency dataset. Note that their time period is different (1982-2002), and they present excess kurtosis.

[6]Data obtained from Yahoo Finance and Kenneth French's Data Library.

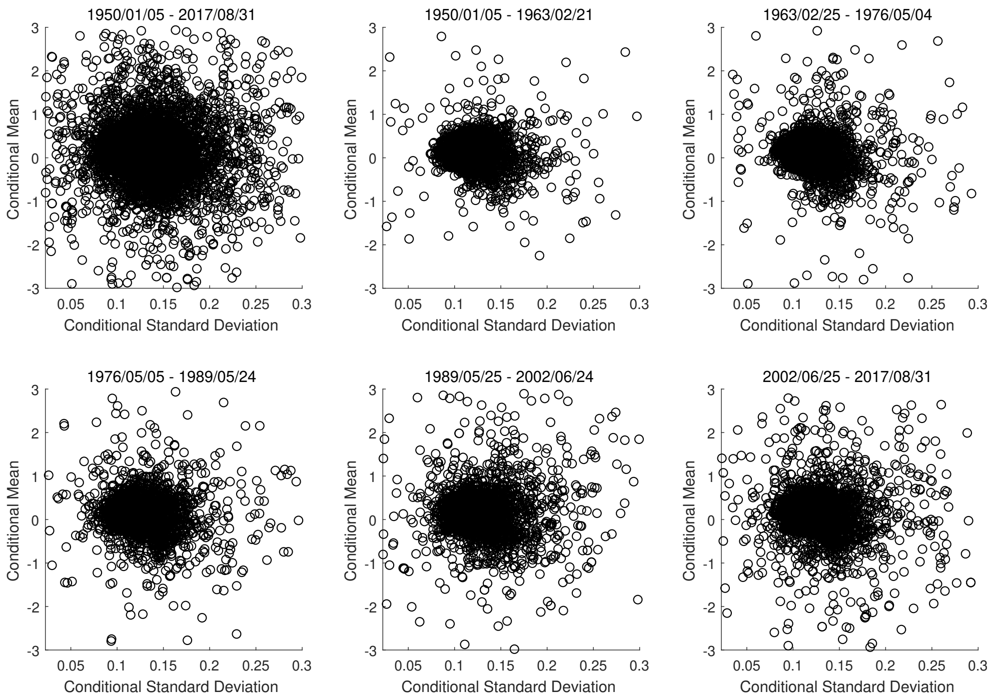


Figure 2.2: Annualized estimates of conditional mean and standard deviation, $p = 4$

Table 2.2: Summary statistics of the estimates $(\widehat{\mu}_t, \widehat{\sigma}_t^2)$

Full Period (1950-2017)

	Mean	Std	Skewness	Kurtosis	AR(1)
$\widehat{\mu}_t$	3.1260×10^{-4}	2.3346×10^{-3}	1.03347	3.4194	0.0190
$\widehat{\sigma}_t^2$	6.5894×10^{-5}	4.8913×10^{-5}	6.6066	76.20624	0.7033

Figure 2.2 reports the (annualized) estimated values, that is, $(\sqrt{252} \cdot \widehat{\sigma}_t, 252 \cdot \widehat{\mu}_t)$, $t = 2, \ldots, n(= 17025)$ when $p = 4$. The result shows there is no noticeable disparity over different time periods, although the estimates are more spreaded out in the more recent periods, showing higher variability. Having a different number of observations does not seem to affect the conclusion either. Interestingly, the number of negative expected excess returns is quite large; such estimates are not inconsistent with asset pricing theory, Boudoukh et al. (1997), Whitelaw (2000), Harvey (2001). The plot of estimates evaluated when $p = 12$ – omitted here – reports similar findings, except that the estimates are a bit more concentrated.

Note that to focus on the main "chunk" of the fitted values, where almost all observations are located, the plots in Figure 2.2 are magnified and truncated on the ranges of $[-3, 3]$ of the y-axis and $[0.0225, 0.3]$ of the x-axis; around 96.1% of the entire fitted values appear in the plot. In particular, among those not appearing in the plot are those with

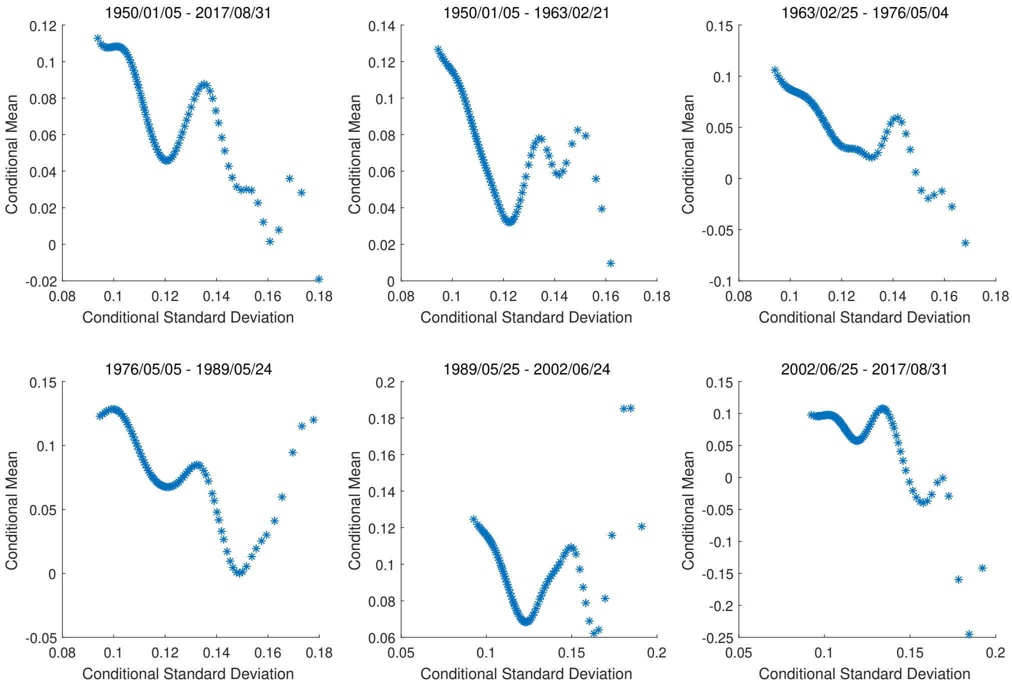


Figure 2.3: Estimated relationship between annualized $\widehat{\sigma}_t$ and $\widehat{\mu}_t$; $p = 4$

zero estimated conditional standard deviation, which constitute around 3% of the whole estimates. This happens when, at a point of evaluation X_t, only one kernel in the sums returns a value of 1 and zero otherwise, so that the second moment equals the squared first moment. One way of reducing the number of such estimates is to increse the bandwidth. We do not proceed to this direction as it will make the bandwidth sub-optimal and currently the number of those observations is still rather negligible.

Figure 2.3 and 2.4 show the estimated relationship obtained using local constant smoothing, with the bandwidth chosen according to Silverman's rule of thumb. The smooths are evaluated at the 100 quantiles of the marginal distribution so that the spacing of the covariate can be shown. The first and last 5 values are taken out, since they tend to be extreme values in general, and including them may make the graph look misleading. All subplots suggest that quadratic fits, i.e. including the conditional variance term, would be appropriate. From Figure 2.4, we note that when $p = 12$, i.e. when the influence of distant lags is "less weighted", we begin to see some negative relationships in some subplots (especially in the more recent periods), which is consistent with our findings later in this section.

Now we consider some parametric analysis, and suppose the conditional mean $\mu(x) = E(Y|X = x)$ and conditional variance $\sigma^2(x) = \mathrm{var}(Y|X = x)$ are related in a quadratic way, i.e.,

$$\mu(x) = \alpha + \beta\sigma(x) + \gamma\sigma^2(x), \tag{2.50}$$

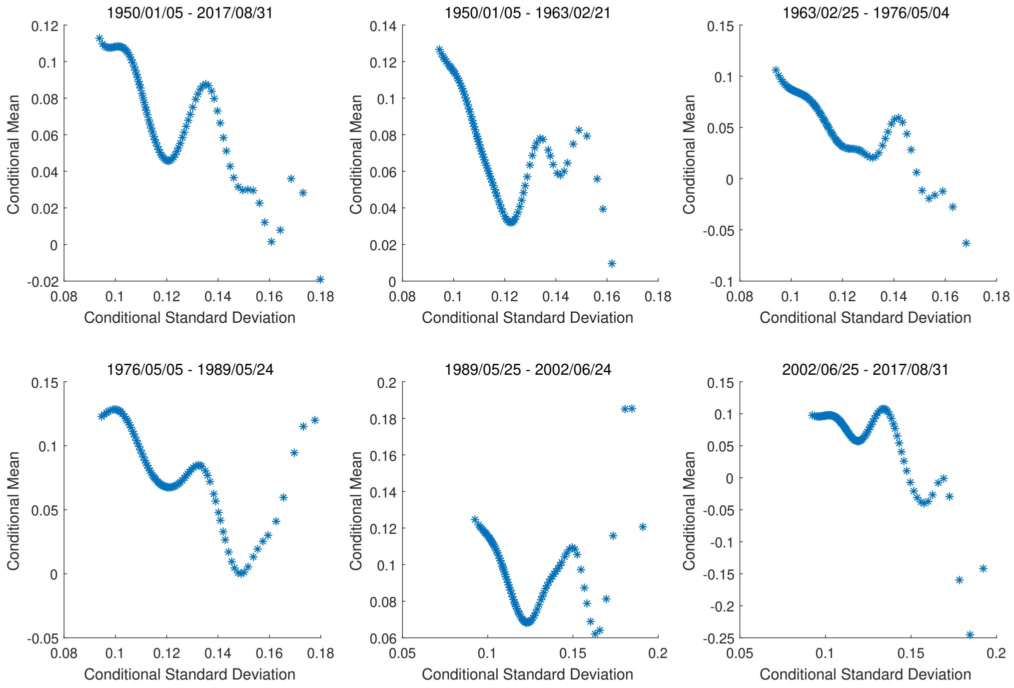


Figure 2.4: Estimated relationship between annualized $\widehat{\sigma}_t$ and $\widehat{\mu}_t$; $p = 12$

where $\theta = (\alpha, \beta, \gamma)^\intercal$ with α, β, γ being unknown constants. Let $x_1, x_2, \ldots, x_q \in \mathbb{R}^\infty$ be some given points such that $\|D^{-1}(x_j - x_k)\| > 0$ for all j, k, and let $\widehat{\mu}(x)$ and $\widehat{\sigma}^2(x)$ be the estimated moments. Then we take

$$\widehat{\theta} = \left(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}\right)^\intercal = \widehat{\Sigma}_q^{-1} \widehat{U}_q$$

$$\widehat{\Sigma}_q = \begin{pmatrix} 1 & \sum_{i=1}^q \widehat{\sigma}(x_i) & \sum_{i=1}^q \widehat{\sigma}^2(x_i) \\ \sum_{i=1}^q \widehat{\sigma}(x_i) & \sum_{i=1}^q \widehat{\sigma}^2(x_i) & \sum_{i=1}^q \widehat{\sigma}^3(x_i) \\ \sum_{i=1}^q \widehat{\sigma}^2(x_i) & \sum_{i=1}^q \widehat{\sigma}^3(x_i) & \sum_{i=1}^q \widehat{\sigma}^4(x_i) \end{pmatrix} \quad ; \quad \widehat{U}_q = \begin{pmatrix} \sum_{i=1}^q \widehat{\mu}(x_i) \\ \sum_{i=1}^q \widehat{\sigma}(x_i)\widehat{\mu}(x_i) \\ \sum_{i=1}^q \widehat{\sigma}^2(x_i)\widehat{\mu}(x_i) \end{pmatrix},$$

where q is finite.

We next derive the limiting distribution of the vector of estimated coefficients $\widehat{\theta} := (\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma})^\intercal$, which can be used for conducting statistical inference. Define:

$$\Sigma_q = \begin{pmatrix} 1 & \sum_{i=1}^q \sigma(x_i) & \sum_{i=1}^q \sigma^2(x_i) \\ \sum_{i=1}^q \sigma(x_i) & \sum_{i=1}^q \sigma^2(x_i) & \sum_{i=1}^q \sigma^3(x_i) \\ \sum_{i=1}^q \sigma^2(x_i) & \sum_{i=1}^q \sigma^3(x_i) & \sum_{i=1}^q \sigma^4(x_i) \end{pmatrix}$$

$$\Omega(x_i) = \begin{pmatrix} \sigma^2(x_i) & \mathrm{skew}(Y_t|X_t = x_i) \\ \mathrm{skew}(Y_t|X_t = x_i) & \sigma^4(x_i)\left(\mathrm{kurt}(Y_t|X_t = x_i) + 2\right) \end{pmatrix} =: \begin{pmatrix} \omega_{1,1}(x_i) & \omega_{1,2}(x_i) \\ \omega_{2,1}(x_i) & \omega_{2,2}(x_i) \end{pmatrix},$$

$$V_q = \sum_{i=1}^{q} J(x_i)\Omega(x_i)J(x_i)^\intercal \quad ; \quad J(x_i) = \begin{pmatrix} 1 & 0 \\ \sigma(x_i) & \frac{\mu}{2\sigma}(x_i) \\ \sigma^2(x_i) & \mu(x_i) \end{pmatrix}.$$

Here, skew and kurt denote skewness and kurtosis of Y_t (conditional on $X_t = x_i$). The result is a direct consequence of consistency of estimated moments and their asymptotic independence across i.

Theorem 2.5. *Let Assumptions B2, B3, B5-B8, and D1-D4 hold, and suppose B4 is strengthened to require* $E(|Y_t|^{8+\delta}) \leq C < \infty$ *for some* $C, \delta > 0$*. Suppose the operator* $g(\cdot) = E(Y^2|X = \cdot)$ *satisfies Assumption B7. Suppose further that* $\omega_{a,b}(u)$ *is continuous in some neighbourhood of* x_i *for all* i*. Then, given the sample observations* $\{Y_t, X_t\}_{t=1}^n$ *specified in A2, we have the following limiting distribution:*

$$\sqrt{nh^{\frac{1-p}{2p-1}} \exp\left(-\kappa_0' h^{-\frac{2}{2p-1}}\right)} \left(\widehat{\theta} - \theta - B_\theta\right) \Longrightarrow N\left(0, \kappa_2(K,p,a)\Sigma_q^{-1} V_q \Sigma_q^{-1}\right),$$

where B_θ *is a bias terms of order* h^β*,* κ_2 *is the constant defined in Section 2.4.2.*

The parameters θ are estimated at the same rate as the functions $\mu(\cdot)$ and $\sigma^2(\cdot)$. It may be possible to achieve faster rates of convergence by allowing $q \to \infty$, as is commonly done in the semiparametric literature, but we have not yet been able to establish this rate improvement; see Chen and Christensen (2015).

With the same S&P500 data as before and the nonparametric estimates we obtained for $p = 4$ and reported in Figure 2.2, we fit the linear regression (2.50). Note that those with zero estimated variance we discussed above are removed (around 3% of whole data), since they can make the fitted estimates misleading and spurious. Also, the standard deviation term is deliberately removed to allow for a direct comparison with the results from those in the existing literature, Pagan and Ullah (1988), Pagan and Hong (1990) and Harvey (2001). We estimate the coefficients α and β, and provide the results along with the values of t-statistics that $\alpha = 0$ and $\beta = 0$ in Table 2.3. The first subperiod is omitted because the estimates for earlier periods may be less reliable due to being evaluated at points with many zeros. Parentheses marked with asterisks (respectively, double asterisks) mean that the corresponding estimates are statistically different from zero at 5% level (respectively, 1% level) of significance based on Newey-West standard errors.

Table 2.3: Estimated parameters obtained using $(252 \cdot \widehat{\mu}_t, 252 \cdot \widehat{\sigma}_t^2)$
Full (1954-2017) and Sub Periods

	Full	1954-1963	1963-1976	1976-1990	1990-2003	2003-2017
α	0.07264	0.10046	0.01672	0.14233	0.09700	0.03761
(t)	$(9.060)^{**}$	$(2.2340)^{*}$	(0.4488)	$(5.0488)^{**}$	$(6.7565)^{**}$	$(2.3038)^{*}$
β	0.40830	-1.64391	2.9360	-3.69919	0.44916	2.14394
(t)	(1.063)	(-0.5213)	(1.2656)	$(-2.9391)^{**}$	(0.6707)	$(2.3265)^{**}$

The result reports a positive effect (0.4083, with $t = 1.063$) of conditional variance on the risk premium during the period of 1954-2017 overall. For the full period we considered a period starting from 1954 here simply because the federal rate series, which we analyse together with later, is available only from 1954. Including the period of 1950-1954 does not change the conclusion. Over the periods of 1963-1976 and 2003-2017 the risk-return relationship is strongly positive, and in the later period the estimate is statistically significant at 1% level. In fact, the estimated risk averse parameter β is 1.41294 (with $t = 2.7317$) on the last two subperiods combined (i.e. the period of around past 30 years), revealing evidence of strongly positive and statistically significant risk-return relation in the recent time after 1990. We may compare these results with the findings of Pagan and Hong (1990, page 61), where they reached a different conclusion with Monthly CRSP data over 1953-1984. They reported the estimated coefficient for conditional variance of -0.87 (with $t = -0.35$). The estimated risk aversion parameter β using our conditional expectations estimates over 1953-1984 is -0.07418 (with $t = -0.0790$), reporting a negative but much weaker risk-return relation.

To investigate how the analysis we adopt in our method may have made any difference, we repeat the same step above by computing nonparametric estimates with using only one lag as conditioning variable. This is to replicate what was done in the papers computed fitted means and variances based on nonparametric regression approaches, e.g. Pagan and Hong (1990). The local constant estimation is done with the standard Gaussian kernel and the bandwidth chosen via cross-validation. We denote by those fitted conditional mean and variance $(\widetilde{\mu}_t, \widetilde{\sigma}_t^2)$, and report the least squares estimates for the parameters α and β below in Table 2.4.

Table 2.4 reveals a very strong and persistent negative risk-return relation throughout all time periods. Over the full period, the estimated risk aversion parameter β is around -3.53, and this is statistically significant at 1% level based on Newey-West standard

Table 2.4: Estimated parameters obtained using $(252 \cdot \widetilde{\mu}_t, 252 \cdot \widetilde{\sigma}_t^2)$
Full (1954-2017) and Sub Periods

	Full	1954-1963	1963-1976	1976-1990	1990-2003	2003-2017
α	0.14977	0.15097	0.15647	0.11693	0.10808	0.19830
(t)	(4.9068)**	(2.0964)*	(4.0584)**	(2.6367)**	(2.2444)**	(2.3897)**
β	-3.33228	-3.14701	-4.30003	-2.38936	-0.77965	-5.13079
(t)	(-2.2834)**	(-0.8642)	(-2.0728)*	(-1.0904)	(-0.3297)	(-1.4274)

errors. This result implies that the conclusion Pagan and Hong (1990) obtained may have been influenced by the fact that they conditioned only on small, fixed lags (when forming the nonparametric estimates). In other words, incorporating further information that are neglected in estimating conditional expectations has clearly led to some new empirical findings. This provides explanations to the conjecture Pagan and Hong (1990) raised in their paper.

2.4.2 Time variation and counter-cyclicity in risk aversion

Meanwhile, the results in Table 3 suggest that the risk-return relationship is strongly time-varying. In particular, over the subperiod 1976-1990 the estimate of β was significantly lower than the other periods. To take a closer look, we conducted a rolling regressions analysis. We set the rolling window to be 4000, roughly a quarter of the number of whole sample, and start estimating β from 1958:09:18. That is, we use conditional expectations estimates over 1958:09:18-1974:12:04 to estimate β for date 1974:12:05, and roll forward the window by one every time. The window size is deliberately chosen to be different from the size of 5 subperiods; this is to check if our previous results in Table 2 are driven by a particular choice of sample size. The time series of estimated parameter β shown below in Figure 5 provides an evidence that investor's average risk aversion has been varying over time.

Furthermore, we observe that interestingly, the time series of risk aversion tends to move in the opposite direction to the federal funds rate[7] f_t, which is a proxy for the business cycle fluctuations, see Figure 5. In fact, the sample correlation between $\widehat{\beta}_t$ and f_t turns out to be -0.5673, implying that the risk aversion exhibits a counter-cyclical behaviour. Also, in Figure 6 we plot the time series of quarterly Sharpe ratio and the designated recession periods by the NBER. The blue line is the ratio computed using our estimates $(\widehat{\mu}, \widehat{\sigma}^2)$, and the red line is the one computed using the standard nonparametric method $(\widetilde{\mu}, \widetilde{\sigma}^2)$. The shadings show that blue line rises over the period of recession in general, which is a finding that is consistent with Lettau and Ludvigson (2010). Note

[7] Data taken from Federal Reserve Bank of St. Louis http://fred.stlouisfed.org

that the red line does not behave as expected in most cases and therefore does not quite capture counter-cyclicity.

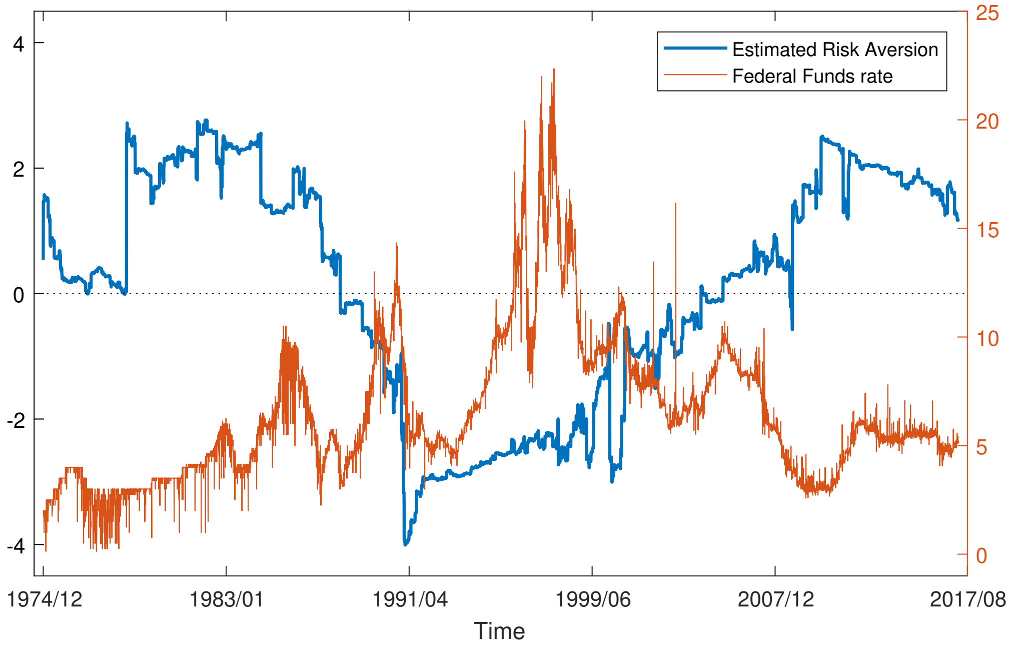


Figure 2.5: Estimated risk-return tradeoff and the federal funds rate

These findings are consistent with what is suggested and widely discussed in the finance literature, for example Antell and Vaihekoski (2016), Campbell and Cochrane (1999), Bliss and Panigirtzoglou (2004), Smith and Whitelaw (2009), Bollerslev, Gibson and Zhou (2011), and Guo, Wang and Yang (2013).

As noted in Mehra (2012), empirical evidence for the financial theory suggesting counter-cyclical risk return tradeoff is rather scarce and limited. Cohn et al. (2015) wrote, "*A key ingredient of many popular asset pricing models is that investors exhibit countercyclical risk aversion, which helps explain major economic puzzles such as the strong and systematic variation in risk premiums over time and the high volatility of asset prices. There is, however, surprisingly little evidence for this ...*"

Our findings on the time series dynamics of risk return tradeoff and their link with the macroeconomy add a supporting empirical evidence to this issue. We reiterate that when standard nonparametric method is employed, these evidence is not well revealed. This potential improvements in the econometric analysis are possibly attributable to the extended flexibility and the inclusion of otherwise neglected information in our method.

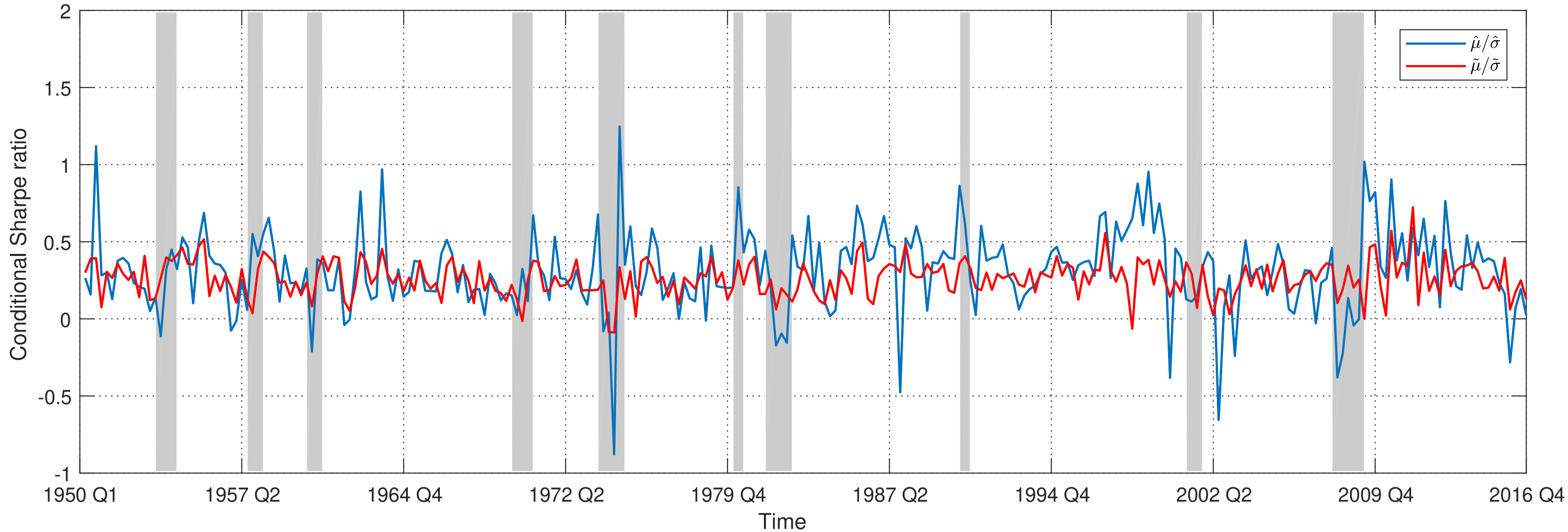


Figure 2.6: Quarterly sharpe ratio (1950-2016) computed based on $(\widehat{\mu}, \widehat{\sigma})$ and $(\widetilde{\mu}, \widetilde{\sigma})$, respectively.
The shades are the recession periods designated by the National Bureau of Economic Research (NBER).

2.4.3 Practical methods for choosing bandwidth

Lastly, we discuss some possible ways for selecting the bandwidth in practice. A natural choice would be direct extensions of two most extensively adopted approaches in the multivariate nonparametric regression, namely, a rule-of-thumb and cross-validation, Green and Silverman (1993), Fan and Gijbels (1996).

We first consider a heuristic plug-in method for obtaining a rule-of-thumb bandwidth. For simplicity, we suppose that the regressors X are i.i.d. Gaussian distributed. This way the small ball probability takes a simple form as discussed in previous sections. Further, this bypasses the need to estimate the $C_{\mathcal{A}}$ term, a function of spectral density representing the degree of dependence between regressors. With the uniform kernel supported on $[0,1]$ i.e. $\lambda = 1$, the asymptotic mean squared error of our estimator is then given by

$$\text{AMSE}(\widehat{m}) = h^{2\beta}\left(\sum_{j=1}^{\infty} c_j j^{p\beta}\right)^2 + \frac{\sigma^2(x)}{nh^{\frac{1-p}{2p-1}}\exp\left(-\kappa_0' h^{-\frac{2}{2p-1}}\right)}. \tag{2.51}$$

Denote by $\mathcal{C}$ the squared term in (2.51), and let $\beta = 1$. Now, differentiating (2.51) with respect to h and equating it to zero we have

$$\begin{aligned}
\frac{\partial\{\text{AMSE}(\widehat{m})\}}{\partial h} &= 2\mathcal{C}h + \frac{\sigma^2(x)}{n(2p-1)}\cdot e^{\kappa_0 h^{-\frac{2}{2p-1}}}\left[(p-1)h^{-\frac{p}{2p-1}} - 2\kappa_0 h^{-\frac{p+2}{2p-1}}\right] = 0\\
\Leftrightarrow \frac{2n(2p-1)\mathcal{C}h}{\sigma^2} &= \exp\left(\kappa_0 h^{-\frac{2}{2p-1}}\right)\cdot\left[2\kappa_0 h^{-\frac{p+2}{2p-1}} - (p-1)h^{-\frac{p}{2p-1}}\right]\\
\Leftrightarrow \frac{14n\cdot\mathcal{C}}{\sigma^2} &= h^{-1}\exp\left(3.605h^{-2/7}\right)\left[7.21h^{-6/7} - 3h^{-4/7}\right] \qquad (2.52)
\end{aligned}$$

where in the last line we substituted $p = 4$ and $\kappa_0 = C^{**} \approx 3.605$ (follows from a straightforward computation; see definitions in Section 2.4.2). As we can substitute the sample variance $\widehat{\sigma}^2$ into σ^2, it now suffices to replace the squared term $\mathcal{C}$ with a suitable estimate.

To proceed, we impose a further restriction and suppose $m(x) = \sum_{j=1}^{\infty}\alpha_j|x_j|$ and $|\alpha_j| \le c_j$ ($= C\theta^j$ for some $0 < \theta < 1$ and constant C[8]). In this case, Assumption B7 in Section 2.4.2:

$$\left|m(x) - m(x')\right| \le \sum_{j=1}^{\infty} c_j\left|x_j - x_j'\right|$$

is satisfied via the reverse triangle inequality. A heuristic idea is to choose some C and θ in such a way that c_j's bound statistically significant estimates of α_j's. Fitting the linear model upto lag 15, say: $Y_t = \sum_{j=1}^{15}\alpha_j|x_j| + \varepsilon_t$, the estimates of α_j at lags $1, 2, 9, 15$ are given by $0.058, 0.0443, 0.0435$ and 0.027, respectively. Therefore, we could let $C = 0.1$

[8] This is a reasonable assumption because $\text{cov}(Y_t, Y_{t-k}) = O(k^{-ck})$ for some c under Assumption A2 and by Davydov's inequality for covariance of mixing sequences.

and $\theta = 0.95$ for example; substituting these values back into the first order condition (2.52) yields

$$\frac{14n}{\widehat{\sigma}^2}\left(\frac{1}{10}\sum_{j=1}^{\infty}(0.95)^j j^4\right)^2 = u^{7/2}e^{3.605u}\Big[7.21u^3 - 3u^2\Big],$$

where $u = h^{-2/7}$, $\widehat{\sigma}^2 \approx 9.3557 \times 10^{-5}$ and $n = 16820$. Numerical approximation via Matlab yields $h = 0.000312$.

An alternative approach would be the cross-validation, where we search for the bandwidth that minimises the mean squared leave-one-out residuals:

$$g(h) := \frac{1}{n}\sum_{t=1}^{n}[Y_t - \widehat{m}_{h,-t}(X_t)]^2, \tag{2.53}$$

where $\widehat{m}_{h,-t}(X_t)$ is the estimate obtained by ignoring t^{th} sample. The result, as illustrated in Figure 2.5, suggests $h = 0.00035$ and $h = 0.00125$ for $p = 4$ and $p = 12$, respectively. These bandwidths are the ones we used earlier in the example in this section. Note that Yao and Tong (1998) proposed an different leave-one-out method for choosing optimal bandwidth for dependent data. When we applied their cross-validation for the data we considered previously however, we noticed that it suggests way lower optimal bandwidths, and the standard approach gives a more reasonable result. This might be because the returns data we consider is almost uncorrelated but nonlinearly dependent.

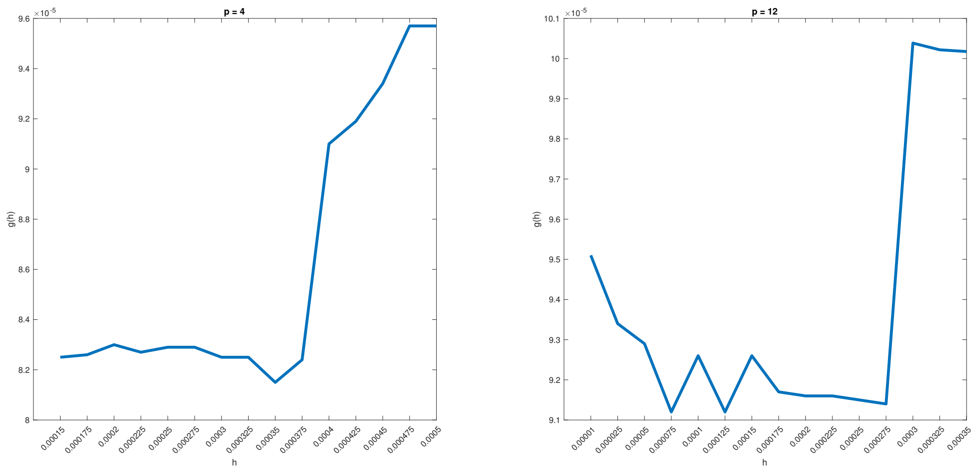


Figure 2.7: Cross validation choice of bandwidth given $p = 4$ and $p = 12$

2.5 Some concluding remarks

In this chapter we studied the nonparametric estimation problem of the infinite order regression. While we answered several open questions raised in the literature, there are some remaining questions that we leave for future studies from a methodological point of view. First, it is not clear how the conclusions we obtained will be changed when the marginal bandwidth is set to decay in a way other than polynomially. It is also a non-trivial question whether the geometric mixing condition in the uniform consistency result could be relaxed to allow weaker dependence of the data. Furthermore, as Linton and Sancetta (2009) pointed out, it may be possible to achieve algebraic convergence rates for some restricted class of functions. For example, we conjecture that given the additive regression model: $\mathrm{E}(Y|X=x) = m(x) = \sum_{j=1}^{\infty} m_j(x_j)$ where $x = (x_j)_j \in \mathbb{R}^{\infty}$, the rate for estimating $m_j(\cdot)$ and $m(\cdot)$ would be the same, just as it was proven to be so in the d-dimensional case (i.e. $m_j(\cdot) = 0$, $\forall j > d \in \mathbb{Z}_+$) by Stone (1985). In this paper, we were unable to find an answer to this question, although we found the existence of the curse of infinite dimensionality under a wide range of frameworks we considered. We leave the question for future study.

Lastly, it would be interesting to come up with a practical way to choose the parameter p or more generally the rate at which the bandwidths expand in the order of the covariates. This should relate to the rate of decay of influence (mixing in the autoregression case) that prevails, and perhaps this can be addressed by using tools from the estimation of memory properties. An alternative approach is to us a penalization method combined with cross-validation, namely, the Bandwidth-LASSO method that adds the penalty $\lambda \sum_i |\phi_i^{-1}|$ to the objective function (2.53). The positive numbers ϕ_i are the bandwidth weights defined previously in Section 2.2.4 and in Assumption B8. The resulting choice of $\{\phi_i^{-1}\}$ will contain many zeros (infinite smoothing of one covariate) depending on the tuning parameter λ, which would give a much more parsimonious representation. The properties of this method will be investigated in the sequel.

Other quantities of interest in prediction such as the conditional median or mode can also be studied. This could be done via nonparametrically estimating the conditional distribution $P(Y \leq y|X = \cdot) = E(1\{-\infty, y](Y)|X = \cdot)$, but would necessarily require a slightly different set of assumptions. It is also quite easy to bring finite dimensional predictors into the theory separately. For example, one may want to allow for slow time variation whereby t/T becomes an additional covariate and the regression function is $m(x, u)$ with $u \in [0, 1]$ and $x \in \mathbb{R}^{\infty}$. In this case we modify the estimator of (2.18) by introducing a multiplicative kernel of the form $k_b(u - t/T)$, where b is a bandwidth and k is a symmetric probability density function.

2.6 Appendix: Proofs of the main results

2.6.1 Proof of Theorem 2.1

PROOF. From the decomposition in (2.21):

$$\widehat{m}(x) - m(x) = \frac{E\widehat{m}_2(x) - m(x)}{\widehat{m}_1(x)} + \frac{\widehat{m}_2(x) - E\widehat{m}_2(x)}{\widehat{m}_1(x)} - \frac{m(x)[\widehat{m}_1(x) - 1]}{\widehat{m}_1(x)},$$

we see that it suffices to show $E\widehat{m}_2(x) - m(x) \to 0$ and $\widehat{m}_2(x) - E\widehat{m}_2(x) \to^P 0$, since $\widehat{m}_1(x) \to^P 1$ will then follow from the latter and complete the proof.

We first consider the "bias component". It is straightforward to see

$$\begin{aligned} E\widehat{m}_2(x) - m(x) &= E\left(\frac{1}{nEK_1}\sum_{t=1}^{n} K_t Y_t - m(x)\right) \\ &= \frac{1}{EK_1}EK_1Y_1 - \frac{EK_1}{EK_1}m(x) = \frac{1}{EK_1}E\left[E\left[\big(Y_1 - m(x)\big)K_1\Big|X\right]\right] \\ &= \frac{1}{EK_1}E\left[\left[m(X) - m(x)\right]K_1\right] \leq \sup_{u\in\mathcal{E}(x,\lambda\underline{h})}\big|m(u) - m(x)\big| \longrightarrow 0 \quad (2.54) \end{aligned}$$

as $n \to \infty$, where K_t is the shorthand notation for $K(\|H^{-1}(x - X_t)\|)$ and $\mathcal{E}(x, \lambda\underline{h})$ is the infinite dimensional hyperellipsoid centred at $x = (x_j)_j \in \mathbb{R}^\infty$ with semi-axes h_j in each direction as introduced in the main text before. The second equality is justified by stationarity that is preserved under measurable transformations, and the last inequality is due to compact support of the kernel and continuity of the regression operator at x (Assumption B1).

The next step concerns with the latter 'variance component' $\widehat{m}_2 - E\widehat{m}_2$. We show its mean-squared convergence to zero. Writing

$$\widehat{m}_2 - E\widehat{m}_2 = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big\{K_tY_t - E(K_tY_t)\Big\} =: \frac{1}{n}\sum_{t=1}^{n}Q_{nt}, \qquad (2.55)$$

we remark that the arguments to follow depend upon the temporal dependence structure of Q_{nt}. In the static regression case, Q_{nt} is a measurable function of $Y_t, X_{1t}, X_{2t}, \ldots$, and hence inherits their joint dependence structure. That is, Q_{nt} is arithmetically α-mixing with the rate specified in A1. In the dynamic regressions case (which covers the autoregression framework), the dependence of Q_{nt} is defined via K_t which is near epoch dependent on (Y_t, V_t) as specified in Assumption A2; this bypasses the issue of Q_{nt} being dependent upon infinite past of Y_t and/or V_t. We proceed with these two cases separately.

CASE 1: STATIC REGRESSION. Clearly, it is sufficient to prove $\mathrm{var}(\widehat{m}_2 - E\widehat{m}_2) \to 0$ for the mean squared convergence. Since Q_{nt} is stationary over time we have

$$\operatorname{var}(\widehat{m}_2 - E\widehat{m}_2) = \frac{1}{n^2}\sum_{t=1}^{n}\operatorname{var}(Q_{nt}) + \frac{2}{n^2}\sum\sum_{1\leq i<j\leq n}\operatorname{cov}\big(Q_{ni}, Q_{nj}\big) \tag{2.56}$$

$$= \frac{1}{n}\operatorname{var}(Q_{n1}) + \frac{2}{n^2}\sum\sum_{1\leq j-i<n}\operatorname{cov}\big(Q_{ni}, Q_{nj}\big)$$

$$= \frac{1}{n}\operatorname{var}(Q_{n1}) + \frac{2}{n^2}\sum_{s=1}^{n-1}(n-s)\cdot\operatorname{cov}\big(Q_{n1}, Q_{n,s+1}\big) =: A_1 + A_2. \tag{2.57}$$

Now, by (2.9), (2.11) and Assumption A it follows that

$$A_1 = \frac{1}{nE^2K_1}\operatorname{var}\bigg(K_1Y_1 - EY_1K_1\bigg) = \frac{\operatorname{var}(K_1Y_1)}{nE^2K_1}$$

$$\leq \frac{EK_1^2Y_1^2}{nE^2K_1} = \frac{E(E(Y_1^2|X_1)K_1^2)}{nE^2K_1} \leq \frac{C}{n\varphi_x(\lambda\underline{h})} \to 0 \quad \text{as} \quad n\to\infty. \tag{2.58}$$

We now move on to the second term A_2 and investigate the covariance term. Since measurable transformations of mixing variables preserve the mixing property, using Davydov's inequality, see Davydov (1968, Lemma 2.1) or Bosq (1996, Corollary 1.1), and stationarity we have

$$\big|\operatorname{cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| = \left|\operatorname{cov}\left(Y_1\frac{K_1}{EK_1}, Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right| \leq \frac{C\{E|Y_1K_1|^{2+\delta}\}^{\frac{2}{2+\delta}}}{\varphi_x(\underline{h}\lambda)^2\cdot s^{k\delta/(2+\delta)}}. \tag{2.59}$$

In the meantime,

$$\big|\operatorname{cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| = \left|\operatorname{cov}\left(Y_1\frac{K_1}{EK_1}, Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right|$$

$$\leq \left|E\left(Y_1\frac{K_1}{EK_1}Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right| + \left|E\left(Y_1\frac{K_1}{EK_1}\right)E\left(Y_{s+1}\frac{K_{s+1}}{EK_1}\right)\right|$$

$$\leq \frac{C}{\varphi_x(\underline{h}\lambda)^2}\left|E\left(K_1K_{s+1}\right)\right| + \frac{C'}{E^2K_1}\left|E\left(K_1\right)E\left(K_{s+1}\right)\right|$$

$$\leq \frac{C}{\varphi_x(\underline{h}\lambda)^2}\cdot\psi_x(\lambda\underline{h}; 1, s+1) + C' \leq C'' \tag{2.60}$$

by stationarity, law of iterated expectation, boundedness of regression function, and Assumption B6, B5 (along with the upper bound $\psi(\lambda\underline{h}; 1, s+1)$ of EK_1K_{s+1} obtained as a direct consequence of B5 following similar arguments used for Lemma 2.1).

With reference to (2.59) and (2.60), we take some increasing sequence $u_n \to \infty$ such that $u_n = o(n)$, and write

$$\sum_{s=1}^{n-1} \big|\mathrm{cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| = \sum_{s=1}^{u_n-1} \big|\mathrm{cov}\big(Q_{n1}, Q_{n,s+1}\big)\big| + \sum_{s=u_n}^{n-1} \big|\mathrm{cov}\big(Q_{n1}, Q_{n,s+1}\big)\big|$$
$$\leq C''\big(u_n - 1\big) + \sum_{s=u_n}^{n-1} \frac{Cs^{-k\delta/(2+\delta)}}{\varphi_x(\underline{h}\lambda)^2} = O\bigg(u_n + \frac{u_n^{-k\delta/(2+\delta)+1}}{\varphi_x(\underline{h}\lambda)^2}\bigg), \tag{2.61}$$

which is $O\big(\varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)}\big)$ upon choosing $u_n \sim \varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)}$.

Consequently, since $k \geq 2(2+\delta)/\delta$ it follows that

$$A_2 := \frac{2}{n^2}\sum_{s=1}^{n-1}(n-s)\cdot \mathrm{cov}\big(Q_{n1}, Q_{n,s+1}\big) = \frac{2}{n}\sum_{s=1}^{n-1}\bigg(1 - \frac{s}{n}\bigg)\cdot \mathrm{cov}\big(Q_{n1}, Q_{n,s+1}\big)$$
$$= O\big(n^{-1}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)} + n^{-2}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)}\big)$$
$$= O\big(n^{-1}[\varphi_x(\underline{h}\lambda)]^{-2(2+\delta)/(k\delta)}\big) = o(1) \tag{2.62}$$

by Assumption B2, and the desired result is obtained.

Case 2: Dynamic Regression.[9] We return back to (2.55):

$$\widehat{m}_2 - E\widehat{m}_2 = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big\{K_tY_t - E(K_tY_t)\Big\} =: \frac{1}{n}\sum_{t=1}^{n} Q_{nt}. \tag{2.63}$$

In this framework $K_t = K(\|H^{-1}(x - X_t)\|)$ is a (measurable) function of $(Y_{t-1}, Y_{t-2}, \ldots)$. Despite loosing the mixing property, K_t inherits stationarity of the mixing process $\{Y_t\}$. We write $K_{t,(r)} = \Psi(Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-r+1}) = E(K_t|Y_t, \ldots, Y_{t-r+1})$ with r as in Assumption A2, and the measurable map Ψ. Then, $K_{t,(r)}$ preserves the mixing dependence structure of Y_t with mixing coefficient $\alpha(\omega - (r-1))$ since $\sigma(K_{s,(r)};\ s \geq t + \omega) \subset \sigma((Y_s, \ldots, Y_{s-r+1});\ s \geq t + \omega) = \sigma(Y_s;\ s \geq t + \omega - (r-1))$.

Now write

$$\widehat{m}_2 - E\widehat{m}_2 = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big[K_{t,(r)}Y_t - E\big(K_{t,(r)}Y_t\big)\Big] + \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big[K_tY_t - K_{t,(r)}Y_t\Big]$$
$$+ \frac{1}{n}\sum_{t=1}^{n}\frac{1}{EK_1}\Big[E\big(K_{t,(r)}Y_t\big) - E(K_tY_t)\Big] = R_1 + R_2 + R_3, \tag{2.64}$$

and first consider the last term R_3. Fix some increasing sequence $q = q_n \to \infty$, and write

[9] For the sake of notational simplicity, we will write the proofs for the dynamic regression framework in terms of its autoregressive special case throughout the appendix. That is, some lags of the response variable Y_t here possibly represent the lagged covariate V_t.

$Y_{t,L} := Y_t 1\{|Y_t| \le q\}$ and $Y_{t,U} = Y_t 1\{|Y_t| > q\}$. Then

$$\begin{aligned} EY_tK_{t,(r)} &= EY_tK\big(\|H^{-1}(x-X_t)\|\big) - EY_{t,U}K\big(\|H^{-1}(x-X_t)\|\big) \\ &\quad + EY_{t,L}K_{t,(r)} - EY_{t,L}K\big(\|H^{-1}(x-X_t)\|\big) \\ &\quad + EY_{t,U}K_{t,(r)} = D_1 + D_2 + D_3. \end{aligned} \tag{2.65}$$

The second part of D_1 is given by

$$\begin{aligned} &EY_{t,U}K\big(\|H^{-1}(x-X_t)\|\big) \le E|Y_t|1_{\{|Y_t|>q\}}K\big(\|H^{-1}(x-X_t)\|\big) \\ &\le q^{-(\delta+1)}E|Y_t|^{2+\delta}1_{\{|Y_t|>q\}}K_t \le Cq^{-(\delta+1)}E|Y_t|^{2+\delta}1_{\{|Y_t|>q\}} = o(q^{-(\delta+1)}) \end{aligned} \tag{2.66}$$

because $1_{\{|Y_t|>q\}} = o(1)$ as $n \to \infty$. Following similar arguments on D_3 we have $D_1 + D_3 = EY_tK_t + o(q^{-(\delta+1)})$. So we are now left with the middle term D_2:

$$D_2 \le E\,|Y_{t,L}|\,\big|K_t - K_{t,(r)}\big| = O\left(q\sqrt{v_2(r_n)}\right) \tag{2.67}$$

by Hölder's inequality. Therefore, from (2.65), (2.66) and (2.67) we see that

$$R_3 = \frac{1}{nEK_1}\sum_{t=1}^{n}\Big[EK_{t,(r)}Y_t - E(K_tY_t)\Big] = o\left(\frac{q^{-(\delta+1)}}{\varphi_x(\lambda\underline{h})}\right) + O\left(\frac{q\sqrt{v_2(r_n)}}{\varphi_x(\lambda\underline{h})}\right), \tag{2.68}$$

and upon choosing $q = (\varphi_x(\underline{h}\lambda)/n)^{-1/(2(\delta+1))}$ we have $o(\varphi^{-1}q^{-(\delta+1)}) = o(\varphi^{-1}(\varphi/n)^{1/2}) = o(n^{-1/2}\varphi^{-1/2}) = o(1)$. Furthermore,

$$\begin{aligned} O\left(\frac{1}{\varphi_x(\underline{h}\lambda)}q\sqrt{v_2(r_n)}\right) &= O\left(\frac{1}{\varphi_x(\underline{h}\lambda)}\cdot\left(\frac{\varphi_x(\underline{h}\lambda)}{n}\right)^{-1/(2(\delta+1))}\sqrt{v_2(r_n)}\right) \\ &= O\left(\frac{\sqrt{v_2(r_n)}}{[\varphi_x(\underline{h}\lambda)]^{(2\delta+3)/(2\delta+2)}n^{-1/(2(\delta+1))}}\right) = o(1) \end{aligned} \tag{2.69}$$

by Assumption A2, yielding $R_3 = o(1)$ and consequently $R_2 = o_p(1)$.

As for the first term that remains,

$$\begin{aligned} R_1 &= \frac{1}{n}\sum_{t=1}^{n}\left[\frac{K_{t,(r)}Y_t - E(K_tY_t)}{EK_1}\right] + \frac{1}{n}\sum_{t=1}^{n}\left[\frac{E(K_tY_t) - E(K_{t,(r)}Y_t)}{EK_1}\right] \\ &= \frac{1}{n}\sum_{t=1}^{n}E\big(Q_{nt}|Y_t, Y_{t-1}, \ldots, Y_{t-r+1}\big) - R_3 \\ &= \frac{1}{n}\sum_{t=1}^{n}Q_{nt,(r)} + o\left(\frac{q^{-(\delta+1)}}{\varphi_x(\underline{h}\lambda)}\right) + O\left(\frac{\sqrt{v_2(r_n)}}{[\varphi_x(\underline{h}\lambda)]^{(2\delta+3)/(2\delta+2)}n^{-1/(2(\delta+1))}}\right). \end{aligned} \tag{2.70}$$

Since $Q_{nt,(r)}$ is α-mixing, we can work with the first term by following similar arguments in the regression case. Specifically, due to boundedness of the kernel and the mixing properties, the bound in (2.59) can be constructed. As for the constant bound constructed in (2.60), we rewrite

$$
\begin{aligned}
\frac{\operatorname{cov}\big(Y_1K_{1,(r)}, Y_{s+1}K_{s+1,(r)}\big)}{\varphi_x(\lambda\underline{h})^2} &= \frac{\operatorname{cov}\big(Y_1[K_{1,(r)}-K_1],\ Y_{s+1}[K_{s+1,(r)}-K_{s+1}]\big)}{\varphi_x(\lambda\underline{h})^2} \\
&\quad + \frac{\operatorname{cov}\big(Y_1[K_{1,(r)}-K_1], Y_{s+1}K_{s+1,(r)}\big)}{\varphi_x(\lambda\underline{h})^2} \\
&\quad + \frac{\operatorname{cov}\big(Y_1, Y_{s+1}[K_{s+1,(r)}-K_{s+1}]\big)}{\varphi_x(\lambda\underline{h})^2} + \frac{\operatorname{cov}\big(Y_1K_1, Y_{s+1}K_{s+1}\big)}{\varphi_x(\lambda\underline{h})^2} \\
&= \mathcal{G}_1+\mathcal{G}_2+\mathcal{G}_3+\mathcal{G}_4.
\end{aligned}
$$

The fourth term $\mathcal{G}_4 \leq C$ by (2.60). Further,

$$
\begin{aligned}
\mathcal{G}_1 \leq &\left|\frac{E(Y_1Y_{s+1}[K_{1,(r)}-K_1][K_{s+1,(r)}-K_{s+1}])}{\varphi_x(\lambda\underline{h})^2}\right| \\
&+\left|\frac{E(Y_1[K_{1,(r)}-K_1])\cdot E(Y_{s+1}[K_{s+1,(r)}-K_{s+1}])}{\varphi_x(\lambda\underline{h})^2}\right| \leq C'\frac{v_2(r)}{\varphi_x(\lambda\underline{h})^2} \to 0
\end{aligned}
$$

by Assumption B6 and by the fact that

$$
\left(\frac{\sqrt{v_2(r_n)}}{\varphi_x(\underline{h}\lambda)}\right) \leq \left(\frac{\sqrt{v_2(r_n)}}{\varphi_x(\underline{h}\lambda)}\right)\cdot (n/\varphi)^{1/(2\delta+2)} \longrightarrow 0
$$

by (2.19) in Assumption A2. Following similar steps it can be shown that $\mathcal{G}_2$ and $\mathcal{G}_3$ converge to zero.

Now choosing an increasing sequence $u_n \sim [\varphi_x(\underline{h}\lambda)^{-2(2+\delta)/(k\delta)} + r_n] \to \infty$ such that $r_n/u_n = o(1)$, we see that (ignoring the array notation in $Q_{nt,(r)}$ for simplicity)

$$
\begin{aligned}
\sum_{s=1}^{n-1}\left|\operatorname{cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\right| &= \sum_{s=1}^{u_n-1}\left|\operatorname{cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\right| + \sum_{s=u_n}^{n-1}\left|\operatorname{cov}\big(Q_{1,(r)}, Q_{s+1,(r)}\big)\right| \\
&\leq C\Big(\varphi_x(\underline{h}\lambda)^{-\frac{2(2+\delta)}{(k\delta)}} + r_n\Big) + \sum_{s=u_n}^{n-1}\frac{C(s-r_n+1)^{-k\delta/(2+\delta)}}{\varphi_x(\underline{h}\lambda)^2} = O\Big(\varphi_x(\underline{h}\lambda)^{-\frac{2(2+\delta)}{(k\delta)}}\Big),
\end{aligned}
$$

since the mixing coefficient for $Q_{nt,(r)}$ denoted $\alpha'(n)$ is given by $\alpha(n-(r-1))$ for $n \geq r$. It now follows by the same arguments in (2.62) that the first term in (2.70) converges to zero, yielding $R_1 = o_p(1)$, which is the result we desired. ■

2.6.2 Proof of Theorem 2.2 and 2.3

PROOF OF THEOREM 2.2 AND 2.3. We start by recalling the bias component discussed in (2.54).

Additional assumptions B7, B8 and D3 allow us to proceed further as follows:

$$\begin{aligned}
\mathcal{B}_n(x) &= E\widehat{m}_2(x) - m(x) = E\left(\frac{1}{nEK_1}\sum_{t=1}^{n} K_t Y_t - m(x)\right) \\
&= \frac{1}{EK_1} EK_1Y_1 - \frac{EK_1}{EK_1} m(x) = \frac{1}{EK_1} E\Big[E\Big[\big(Y_1 - m(x)\big)K_1 \Big| X\Big]\Big] \\
&= \frac{1}{EK_1} E\Big[\Big[m(X) - m(x)\Big]K_1\Big] \le \sup_{u\in\mathcal{E}(x,\lambda\underline{h})} \big|m(u) - m(x)\big| \\
&\le \sup_{u\in\mathcal{E}(x,\lambda\underline{h})} \sum_{j=1}^{\infty} c_j \big|u_j - x_j\big|^{\beta} = \sum_{j=1}^{\infty} c_j(\lambda h \phi_j)^{\beta} = h^{\beta}\left(\lambda^{\beta}\sum_{j=1}^{\infty} c_j j^{p\beta}\right) < \infty. \qquad (2.71)
\end{aligned}$$

Now rewriting the decomposition (2.21) as

$$\begin{aligned}
&\widehat{m}(x) - m(x) - \mathcal{B}_n(x) \\
&\qquad = \frac{\mathcal{B}_n(x)\cdot\big[1 - \widehat{m}_1(x)\big]}{\widehat{m}_1(x)} + \frac{\widehat{m}_2(x) - E\widehat{m}_2(x) - m(x)\big[\widehat{m}_1(x) - 1\big]}{\widehat{m}_1(x)},
\end{aligned}$$

and noting that $\widehat{m}_1(x) \to^p 1$ (an immediate consequence of Theorem 2.1), we see that it suffices to derive the limiting distribution of

$$\begin{aligned}
&\widehat{m}_2(x) - E\widehat{m}_2(x) - m(x)[\widehat{m}_1(x) - 1] \\
&= \frac{1}{n}\sum_{t=1}^{n} \frac{1}{EK_1}\Big[K_tY_t - m(x)K_t - E(K_tY_t) + m(x)EK_t\Big] =: \frac{1}{n}\sum_{t=1}^{n} R_{nt}. \qquad (2.72)
\end{aligned}$$

In the rest of the proof, the way how we construct the general CLT *under Assumption A1* is quite similar to the proofs of theorems in Masry (2005), where asymptotic normality is established in a functional context for mixing data sample. For completeness of the proof however, we will go over some of the main arguments; some relatively less important details will only be briefly skeched to prevent being repetitive.

By Assumption B6, D3, D4, and the law of iterated expectations, the asymptotic variance of the triangular array R_{nt} is given by

$$\begin{aligned}
\mathrm{var}(R_{nt}) &= \frac{\mathrm{var}[K_t(Y_t - m(x))]}{E^2K_1} = \frac{1}{E^2K_1}\left\{E\Big[K_t(Y_t - m(x))\Big]^2 - E^2\big[K_t(Y_t - m(x))\big]\right\} \\
&\simeq \frac{1}{E^2K_1}\left\{E\big[\sigma^2(X)K_1^2\big] + E\left(\Big[m(X) - m(x)\Big]^2 K_1^2\right)\right\} \\
&= \frac{1}{E^2K_1}\left\{\sigma^2(x)EK_1^2 + E\left(\Big[\sigma^2(X) - \sigma^2(x)\Big]K_1^2\right) + o(1)EK_1^2\right\} \\
&= \frac{EK_1^2}{E^2K_1}(\sigma^2(x) + o(1)) \simeq \frac{\sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2}. \qquad (2.73)
\end{aligned}$$

Using the latter assumption of D4 and Assumptions B, and following similar arguments as in the above and those in the proof of Theorem 2.1, it can be readily shown that the covariance term is of negligible order, which together with (2.73) shows (2.31).

Meanwhile, under Assumption D1 the small ball probability can be written in terms of the centered small deviation and $p^*(\cdot)$, the Radon-Nikodym derivative of the induced probability measure P_{z-Z} with respect to P_Z:

$$\begin{aligned}\varphi_x(\lambda\underline{h}) &= P\big(X \in \mathcal{E}(x, \lambda\underline{h})\big)\\ &= P\bigg(\sum_{j=1}^{\infty} j^{-2p}\big(x_j - X_j\big)^2 \leq h^2\lambda^2\bigg) = P\Big(\|z - Z\| \leq h\lambda\Big)\\ &= \int_{B(0,h\lambda)} dP_{z-Z}(u) = \int_{B(0,h\lambda)} p^*(u)\, dP_Z(u)\\ &\simeq p^*(0) \cdot P(\|Z\| \leq h\lambda) = p^*(0) \times P\left(\sum_{j=1}^{n} j^{-2p} X_j^2 \leq h^2\lambda^2\right). \end{aligned} \tag{2.74}$$

Given that the fourth moment of X_j is finite by Assumptions C, the latter probability in (2.74) can be explicitly specified by substituting $r = h^2\lambda^2$, $A = 2p$, and $a = 2p/(2p-1)$ in Proposition 4.1 of Dunker et al. (1998) for the i.i.d. case. When the marginal regressors are dependent as in Assumption C2, the small ball probability can be specified (by letting $r = h^2\lambda^2 C_{\mathcal{A}}^{-2}$ and leaving the others the same) in view of Theorem 1.1 of Hong, Lifshits and Nazarov (2016). In the general i.i.d. case (under Assumptions A1 and C1) we have

$$\frac{\sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2} = \frac{1}{\phi(h)} \cdot \frac{\sigma^2(x)\xi_2}{p^*(0)\xi_1^2} \cdot \frac{C^* C_\ell}{\lambda^{\frac{1+2\rho p}{2p-1}}},$$

where $\phi(h) = h^{(1+2\rho p)/(2p-1)} \exp\{-C^{**}(\lambda h)^{-2/(2p-1)}\}$ and

$$C_\ell = \lim_{h\to 0}\left[\ell^{-1/2}\left(h^{-\frac{4p}{2p-1}}\right)\right] \qquad C^* = \frac{(2\pi)^{(1+2p\rho)}(2p-1)}{\Gamma^{-1}(1-\rho)\cdot(2p)^{\frac{2p(\rho+2)-1}{2p-1}}} \cdot \zeta^{\frac{2p(1+\rho)}{2p-1}}.$$

$\Gamma(\cdot)$ is the Gamma function, ξ_1 and ξ_2 are the constants specified in (2.12), and λ is the upper bound of the support of the kernel. The constants for the dependent Gaussian case (that is, under Assumption C2) can be specified similarly.

In constructing the central limit theorem we consider the normalized statistic $R_{nt}^* := \sqrt{\phi(h)} \cdot R_{nt}$ and derive the limiting distribution of $(1/\sqrt{n}) \cdot R_{nt}^*$. We shall prove under Assumption A2 as it involves some further arguments, without which the proof just serves as the proof under Assumption A1. We make use of the standard Bernstein's blocking method and partition $\{1, \ldots, n\}$ by $2k(= 2k_n \to \infty)$ number of blocks of two different sizes that alternate (hereafter referred to as the "big" and "small" blocks) and lastly a single block (the "last block") that covers the remainder. The size of the alternating blocks is given by a_n and b_n respectively, where the one for the "big-blocks" a_n is set to dominate that for the "small-blocks" b_n in large sample, i.e. $b_n = o(a_n)$. Specifically, take $k_n = \lfloor n/(a_n + b_n) \rfloor$ and $a_n = \lfloor \sqrt{n\phi(h)}/q_n \rfloor$, where $q_n \to \infty$ is a sequence of integer; it then clearly follows that $a_n/n \to 0$ and $a_n/\sqrt{n\phi(h)} \to 0$. We also assume $(n/a_n) \cdot \alpha^*(b_n) = (n/a_n) \cdot \alpha(b_n - r + 1) \to 0$, where α^* is the mixing coefficient of $R^*_{nt,(r)} = E(R^*_{nt}|\mathcal{F}^{t-1}_{t-r+1})$.

By construction above we can write $\sqrt{n}^{-1} \sum_{t=1}^n R^*_{nt}$ as the sum of the groups of big-blocks $\mathcal{B}$, small-blocks $\mathcal{S}$ and the remainder block $\mathcal{R}$ defined as

$$\begin{aligned}
\mathcal{B} &:= \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \Xi_{1,j} = \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R^*_{nt} \right) \\
\mathcal{S} &:= \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \Xi_{2,j} = \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R^*_{nt} \right) \\
\mathcal{R} &:= \frac{1}{\sqrt{n}} \Xi_{3,j} = \frac{1}{\sqrt{n}} \left(\sum_{t=k(a+b)+1}^{n} R^*_{nt} \right).
\end{aligned}$$

The aim is to show that the contributions from the small and the last remaining block are negligible, and that the big-blocks are asymptotically independent. Consider the big blocks $\mathcal{B}$. Given r as in Assumption A2, and $R^*_{nt,(r)} = E(R^*_{nt}|Y_t, \ldots, Y_{t-r+1})$ we have

$$\mathcal{B} = \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R^*_{nt,(r)} \right) + \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \left(\sum_{t=j(a+b)+1}^{j(a+b)+a} \left[R^*_{nt,(r)} - R^*_{nt} \right] \right) = \mathcal{Q}_1 + \mathcal{Q}_2.$$

As for the second term, consider

$$\begin{aligned}
\frac{1}{\sqrt{n}} E\mathcal{Q}_2 &\leq \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \sum_{t=j(a+b)+1}^{j(a+b)+a} E\left| R^*_{nt,(r)} - R^*_{nt} \right| \\
&= \frac{1}{EK_1} \frac{1}{\sqrt{n}} \sum_{j=0}^{k-1} \sum_{t=j(a+b)+1}^{j(a+b)+a} E\left| K_t Y_t - Y_t E(K_t | Y_t, Y_{t-1}, \ldots, Y_{t-r+1}) \right| \\
&\leq \frac{1}{\sqrt{n}} \frac{1}{\varphi_x(\underline{h}\lambda)} \sum_{j=0}^{k-1} \sum_{t=j(a+b)+1}^{j(a+b)+a} E|Y_t||K_t - K_{t,(r)}| \\
&\leq \frac{1}{\sqrt{n}} \frac{1}{\varphi_x(\underline{h}\lambda)} \sum_{j=0}^{k-1} \sum_{t=j(a+b)+1}^{j(a+b)+a} \left(E|Y_t|^2 \right)^{1/2} \left(E|K_t - K_{t,(r)}|^2 \right)^{1/2} \\
&\leq C \cdot \frac{1}{\sqrt{n}} k_n a_n \frac{\sqrt{v_2(r_n)}}{\varphi_x(\lambda \underline{h})} = O\left(\frac{\sqrt{n \cdot v_2(r_n)}}{\varphi_x(\lambda \underline{h})} \right) = o(1),
\end{aligned}$$

which implies that $\sqrt{n}^{-1}\mathcal{Q}_2 = o_p(1)$.

We now show asymptotic independence of terms in $\mathcal{Q}_1$, on noting that $\Xi'_{1,j}s$ are independent if for all real t_j

$$\left| E\left[\sum_{j=0}^{k-1} \exp\left(\mathrm{i}t_j \Xi_{1,j}\right)\right] - \prod_{j=0}^{k-1} E\left[\exp(\mathrm{i}t_j \Xi_{1,j})\right] \right| \tag{2.75}$$

is zero. Applying the Volkonskii-Rozanov inequality (see Fan and Yao (2003, page 72) for example), it can be shown that (2.75) is bounded above by $C(n/a_n) \cdot \alpha(b_n - r + 1) \to 0$, implying asymptotic independence.

Moving on to the small blocks, due to stationarity we have

$$\begin{aligned}
\operatorname{var}(\mathcal{S}) &= \frac{1}{n}\operatorname{var}\left(\sum_{j=0}^{k-1}\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R^*_{nt}\right) \\
&= \frac{1}{n}\sum_{j=0}^{k-1}\operatorname{var}\left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R^*_{nt}\right) + \frac{1}{n}\sum_{j\neq l}^{k-1}\sum \operatorname{cov}\left(\sum_{t=j(a+b)+a+1}^{(j+1)(a+b)} R^*_{nt}, \sum_{s=l(a+b)+a+1}^{(l+1)(a+b)} R^*_{ns}\right) \\
&= \frac{1}{n}\sum_{j=0}^{k-1}\left(b_n \operatorname{var}(R^*_{nt}) + \sum_{t\neq l}^{b_n}\operatorname{cov}(R^*_{nt}, R^*_{nl})\right) + \frac{1}{n}\sum_{j\neq l}^{k-1}\sum\sum_{i,j=1}^{b_n}\operatorname{cov}\big(R^*_{n,i+w_j}, R^*_{n,r+w_l}\big) \\
&= Q_1 + Q_2 + Q_3.
\end{aligned}$$

where $w_j = j(a+b) + a$.

Regarding the first term, similar arguments used in deriving (2.73) yield

$$Q_1 = \frac{1}{n}k_n b_n \frac{\left[\varphi_x(\underline{h}\lambda)^{1/2}\right]^2 \sigma^2(x)\xi_2}{\varphi_x(\underline{h}\lambda)\xi_1^2} = \frac{k_n b_n \sigma^2(x)\xi_2}{n\xi_1^2} \longrightarrow 0 \tag{2.76}$$

because $k_n b_n/n \sim b_n/(a_n+b_n) \to 0$. Now moving on to Q_2 and Q_3, the sum of covariances can be dealt with in the same manner as we did for the variance using (2.73), so $Q_2 \to 0$. Similarly for Q_3, implying $\operatorname{var}(\mathcal{S}) \to 0$ as desired. Convergence result for the remainder $\mathcal{R}$ can be established similarly, and is bounded by $C(a_n + b_n)/n \to 0$.

The results above suggest that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} R^*_{nt} = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a} R^*_{nt}\right) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j + o_p(1), \tag{2.77}$$

and the desired result holds in view of (62) and the CLT for triangular array upon checking the Lindeberg condition (which is omitted here due to its similarity with Masry (2005, page 174-175)). Corollary 2.2 now follows because

$$\sqrt{n\phi(h)}\left(\frac{\widehat{m}-m-\mathcal{B}_n}{\sqrt{n\phi(h)}\Delta_n}\right)=\frac{\sqrt{n}\frac{1}{n}\sum_{t=1}^{n}R_{nt}^{*}}{\sqrt{\frac{1}{n}\sum_t \widehat{R}_{nt}^{*,2}}}=\frac{\frac{1}{\sqrt{n}}\sum_{t=1}^{n}R_{nt}^{*}}{\sqrt{\frac{1}{n}\sum_t R_{nt}^{*,2}+o_p(1)}}$$

$$=\frac{\frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\sum_{t=j(a+b)+1}^{j(a+b)+a}R_{nt}^{*}+o_p(1)}{\sqrt{\frac{1}{n}\sum_{j=0}^{k-1}\left(\sum_{t=j(a+b)+1}^{j(a+b)+a}R_{nt}^{*}\right)^2+o_p(1)}}=\frac{\frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j+o_p(1)}{\sqrt{\frac{1}{n}\sum_{j=0}^{k-1}\eta_j^2+o_p(1)}}\Longrightarrow N(0,1) \qquad (2.78)$$

by Theorem 4.1 of de la Peña et al. (2009), since the denominator converges in probability to a strictly positive quantity ($\sigma^2(x)\xi_2/\xi_1^{-2}$), and that η_j belongs to the domain of attraction of a normal distribution by definition and (2.77). ■

2.6.3 Proof of Lemmas 2.1 and 2.2

Proof. Lemma 2.1 is a straightforward extension of Lemma 4.3 and 4.4 of Ferraty and Vieu (2006), and hence is omitted. Lemma 2.2 can be shown by noting that for each n the τ_n-dimensional polyhedron $D := \{w = (w_i)_{i\leq\tau} \in \mathbb{R}^\tau, |w_i| \leq \lambda\}$ can be covered by $([2\lambda\sqrt{\tau}/\varepsilon + 1])^\tau$ number of balls of radius ε, see Chaté and Courbage (1997), and then following the arguments of the proof of Theorem 2 in Jia et al. (2003). ■

2.6.4 Proof of Theorem 2.4 and 2.5

Proof of Theorem 2.4. In the sequel, we omit the subscript τ in the notations for truncated regressor and its estimator, i.e. $m_\tau(\cdot)$ and $\widehat{m}_\tau(\cdot)$ for notational simplicity. As before, we start from the decomposition (2.21):

$$\widehat{m}(x)-m(x)=\frac{1}{\widehat{m}_1(x)}\left(\left[\widehat{m}_2(x)-E\widehat{m}_2(x)\right]+\left[E\widehat{m}_2(x)-m(x)\right]-m(x)\left[\widehat{m}_1(x)-1\right]\right).$$

We recall from (2.74) that $\varphi_x(\lambda\underline{h}) \sim \varphi(\lambda\underline{h})$. Further, notice that the small deviation for the truncated regressor $X = (X_1, \ldots, X_\tau, 0, 0, \ldots)$ denoted $\varphi^{\mathcal{T}}(\lambda\underline{h})$ satisfies

$$\varphi(\lambda\underline{h})=P\left(\sum_{j=1}^{\infty}j^{-2p}X_j^2\leq h^2\right)\leq P\left(\sum_{j=1}^{\tau}j^{-2p}X_j^2\leq h^2\right)=\varphi^{\mathcal{T}}(\lambda\underline{h}). \qquad (2.79)$$

Note that as implicitly mentioned in the main text, (2.43) is meant to hold for $\varphi^{\mathcal{T}}(\lambda\underline{h})$.

In the first step of the proof we show

$$\sup_{x\in\mathcal{S}_\tau}\left|\widehat{m}_2(x)-E\widehat{m}_2(x)\right|=O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right). \qquad (2.80)$$

We cover the set $\mathcal{S}_\tau$ defined in (2.41) with $L = L(\mathcal{S}_\tau, \eta)$ number of balls of radius η denoted by I_k, each of which is centred at x_k, $k = 1, \ldots, L$. i.e. $S_\tau \subset \bigcup_{k=1}^{L_n} B(x_k, \eta)$. Then it follows that

$$\begin{aligned}
\sup_{x\in\mathcal{S}_\tau}\Big|\widehat{m}_2(x) - E\widehat{m}_2(x)\Big| &= \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\Big|\widehat{m}_2(x) - E\widehat{m}_2(x)\Big| \\
&= \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\Big|\widehat{m}_2(x) - \widehat{m}_2(x_k) + \widehat{m}_2(x_k) - E\widehat{m}_2(x_k) + E\widehat{m}_2(x_k) - E\widehat{m}_2(x)\Big| \\
&\le \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\Big|\widehat{m}_2(x) - \widehat{m}_2(x_k)\Big| + \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\Big|E\widehat{m}_2(x_k) - E\widehat{m}_2(x)\Big| \\
&\quad + \max_{1\le k\le L_n}\Big|\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)\Big| \;=: R_1 + R_2 + R_3, \qquad (2.81)
\end{aligned}$$

where $\widehat{m}_2(x_k) = (nEK_1)^{-1}\sum_{t=1}^n Y_t K_{t,k}$ and $K_{t,k} = K(\|H^{-1}(x_k - X_t)\|)$.

We first consider R_1. By Lemma 2.1,

$$\begin{aligned}
R_1 &= \max_{1\le k\le L_n}\sup_{x\in I_k\cap S_\tau}\Big|\widehat{m}_2(x) - \widehat{m}_2(x_k)\Big| \\
&= \max_{1\le k\le L_n}\sup_{x\in I_k\cap S_\tau}\left|\frac{1}{nEK_1}\sum_{t=1}^n Y_t K\Big(\|H^{-1}(x - X_t)\|\Big) - Y_t K\Big(\|H^{-1}(x_k - X_t)\|\Big)\right| \\
&\le \max_{1\le k\le L_n}\sup_{x\in I_k\cap S_\tau}\frac{C}{n\varphi^{\mathcal{T}}(\lambda\underline{h})}\sum_{t=1}^n \Big|Y_t K_t - Y_t K_{t,k}\Big| \cdot 1_{\mathcal{E}(x,\lambda\underline{h})\cup\mathcal{E}(x_k,\lambda\underline{h})}(X_t).
\end{aligned}$$

Now because type-I kernels are Lipschitz continuous on $[0,\lambda]$, by the triangle inequality we have

$$R_1 \le \frac{1}{n}\sum_{t=1}^n \frac{C'|Y_t|}{\varphi^{\mathcal{T}}(\underline{h}\lambda)}\eta h^{-1}\cdot 1_{\mathcal{E}(x,\lambda\underline{h})\cup\mathcal{E}(x_k,\lambda\underline{h})}(X_t) =: \frac{1}{n}\sum_{t=1}^n J_t,$$

where J_t is α-mixing under both assumptions A1$'$ and A2$'$ (with a different rate under A2$'$: $\alpha^*(n) = \alpha(n - \tau + 1)$, where $\alpha(\cdot)$ is the mixing rate under A1$'$). Let $\eta = \log n/n^2$. Using Assumption B6 and the law of iterated expectations it is straightforward to see that

$$E|J_t| \le \frac{C\eta}{h}. \qquad (2.82)$$

Using Lemma 2.2 we can specify the Kolmogorov's entropy of S_τ for $\eta = \log n/n^2$:

$$\log L\left(S, \frac{\log n}{n^2}\right) = C\log\left[\left(\frac{2\lambda n^2}{\sqrt{\log n}} + 1\right)^{\log n}\right] \sim \log n \times \log\left[\frac{2\lambda n^2}{\sqrt{\log n}}\right],$$

implying that the order of Kolmogorov's $\frac{\log n}{n^2}$ entropy is of order $(\log n)^2$.[10]

We now apply the Fuk-Nagaev inequality (see for example, Fuk and Nagaev (1971),

[10] Notice that in this case (2.43) is indeed satisfied with $\beta = 1, p = 4, \epsilon = 1/4$, for example.

or Rio (2000)) for exponentially mixing variables in Merlevède, Peligrad and Rio (2011, 1.7) with $\varepsilon = \varepsilon_0[\log L\big(S, \frac{\log n}{n^2}\big)/(n\varphi(\lambda\underline{h}))]^{1/2}$ and $r = (\log L)/\varphi(\lambda\underline{h})$, where ε_0 is some positive constant. Since

$$s_n^2 := \sum_{t=1}^{n}\sum_{s=1}^{n} \operatorname{cov}(J_t, J_s) = O\big(n\varphi^{\mathcal{T}}(\lambda\underline{h})^{-1}\log n\big)$$

and the required tail condition holds, under Assumption A1′ we obtain

$$\begin{aligned}
P\left(\left|\frac{1}{n}\sum_{t=1}^{n} J_t - EJ_t\right| > \varepsilon\right) &= P\left(\left|\sum_{t=1}^{n}(J_t - EJ_t)\right| > n\varepsilon_0\sqrt{\frac{\log L\big(S, \frac{\log n}{n^2}\big)}{n\varphi(\lambda\underline{h})}}\right) \\
&\le\ 4\left(1 + \frac{n^2\varepsilon_0^2 \log L\big(S, \frac{\log n}{n^2}\big)}{16 r s_n^2 n\varphi(\lambda\underline{h})}\right)^{-\frac{r}{2}} + \frac{16C\sqrt{n\varphi(\lambda\underline{h})}}{\varepsilon_0\sqrt{\log L}} \exp\left(-\varsigma\left[\frac{\frac{1}{4}n\varepsilon_0\sqrt{\frac{\log L}{n\varphi(\lambda\underline{h})}}}{\log L/\varphi(\lambda\underline{h})}\right]^{\gamma}\right) \\
&\le\ 4\left(1 + \frac{C\varepsilon_0^2\log n}{16r}\right)^{-\frac{r}{2}} + \frac{16C\sqrt{n\varphi(\lambda\underline{h})}}{\varepsilon_0\log n}\exp\left(-\varsigma\left[\frac{\varepsilon_0\sqrt{n\varphi(\lambda\underline{h})}}{4\log n}\right]^{\gamma}\right) \\
&\le\ 4\left(1 + \frac{C\varepsilon_0^2\varphi(\lambda\underline{h})}{16\log n}\right)^{-\frac{\log L}{2\varphi(\lambda\underline{h})}} + \frac{16C}{\varepsilon_0}\left(\frac{\sqrt{n\varphi(\lambda\underline{h})}}{\log n}\right)\exp\left(-\varsigma\varepsilon_0^{\gamma}4^{-\gamma}\cdot\left[\frac{\sqrt{n\varphi(\lambda\underline{h})}}{\log n}\right]^{\gamma}\right) \\
&\le\ 4\exp\left(-\frac{\varepsilon_0^2 C\log n}{32}\right) + \frac{16C}{\varepsilon_0}\left(\frac{\sqrt{n\varphi(\lambda\underline{h})}}{\log n}\right)\cdot e^{-C'(\sqrt{n\varphi}/\log n)} \longrightarrow 0, \qquad (2.83)
\end{aligned}$$

where $\varsigma > 1$ and $\gamma \geq 1$ are as defined in Section 2.4.4, by choosing ε_0 sufficiently large. In the last inequality we exploited the fact that $\log(1+\epsilon) = \epsilon + o(\epsilon^2)$ as $\epsilon \to 0$. Under Assumption A2′, a penalty of $(-\log n)$ is incurred in the squared brackets in the inequalities above. This does not affect the conclusion (2.83) because $\tau = \log n \leq (\log n)^2 \leq \sqrt{n\varphi}/(\log n)^{1+\epsilon} \leq \sqrt{n\varphi}/\log n$ by (2.43) in Assumption E.[11]

Therefore, in view of (2.82) it now follows that

$$\begin{aligned}
R_1 = \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau}\left|\widehat{m}_2(x) - \widehat{m}_2(x_k)\right| &\le O\left(\frac{\eta}{h}\right) + O_P\left(\sqrt{\frac{\log L\big(S, \frac{\log n}{n^2}\big)}{n\varphi(\lambda\underline{h})}}\right) \\
&= O\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right) + O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right) = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right). \qquad (2.84)
\end{aligned}$$

As for the second term R_2, we have

$$R_2 \le \max_{1\le k\le L_n}\sup_{x\in I_k\cap\mathcal{S}_\tau} E\left|\widehat{m}_2(x) - \widehat{m}_2(x_k)\right| = O\left(\frac{\eta}{h}\right) = O\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda\underline{h})}}\right). \qquad (2.85)$$

[11]To elaborate, this is due to the fact that $y\exp(-(y - g(y))) \to 0$ as $y \to \infty$, as long as $(y - g(y))$ tends to $+\infty$ as $y \to \infty$ at the speed strictly faster than $\log y$.

Next we move on to the last component:

$$R_3 = \max_{1\le k\le L_n} \left|\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)\right| =: \max_{1\le k\le L_n} \left|W_n(x_k)\right| \tag{2.86}$$

where

$$\begin{aligned} W_n(x) = \widehat{m}_2(x) - E\widehat{m}_2(x) &= \frac{1}{nEK_1}\sum_{t=1}^{n}\Big[Y_tK_t - EY_tK_t\Big] \\ &\le \frac{C}{n\varphi^{\mathcal{T}}(\underline{h}\lambda)}\sum_{t=1}^{n}\Big[Y_tK_t - EY_tK_t\Big] = \frac{1}{n}\sum_{t=1}^{n}U_{nt} \end{aligned}$$

where $U_{nt} = (\varphi^{\mathcal{T}}(\underline{h}\lambda))^{-1}C(Y_tK_t - EY_tK_t)$.

By following similar arguments in the proof of Theorem 2.1, it can be readily seen that

$$s_n^2 = \sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{cov}\left(U_{nt}, U_{ns}\right) = O\left(n\varphi^{\mathcal{T}}(\underline{h}\lambda)^{-1}\right).$$

With the exponential tail condition in B4, we apply the same Fuk-Nagaev inequality for exponentially mixing sequences we referred to in the above. Writing $L_n := L\left(S, \frac{\log n}{n^2}\right)$ and taking $\varepsilon = \varepsilon_0[\log L\big(S, \frac{\log n}{n^2}\big)/(n\varphi(\lambda\underline{h}))]^{1/2}$ and $r = (\log n)^{2+\epsilon}/\varphi^{\mathcal{T}}(\lambda\underline{h})$, $\epsilon \in (0, 1/2)$ for some $\varepsilon_0 > 0$, under Assumption A1$'$ we have

$$\begin{aligned} &P\left(\max_{1\le k\le L_n} |\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)| > \varepsilon\right) \\ &\quad\le\ L_n\cdot\sup_{x\in\mathcal{S}} P\left(|W_n(x)| > \varepsilon_0\sqrt{\frac{\log L_n}{n\varphi(\lambda\underline{h})}}\right) \\ &\quad\le\ L_n\cdot\sup_{x\in\mathcal{S}} P\left(\left|\sum_{t=1}^{n}U_{nt}\right| > n\varepsilon_0\sqrt{\frac{\log L_n}{n\varphi^{\mathcal{T}}(\lambda\underline{h})}}\right) \\ &\quad\le\ L_n\cdot 4\left(1 + \frac{n^2\varepsilon_0^2\log L_n}{16rs_n^2n\varphi^{\mathcal{T}}(\lambda\underline{h})}\right)^{-\frac{r}{2}} \\ &\qquad\quad + \frac{16L_nCn\sqrt{n\varphi^{\mathcal{T}}(\lambda\underline{h})}}{n\varepsilon_0\log n}\exp\left(-\varsigma\left\{\frac{\varepsilon_0\sqrt{n}\log n/\sqrt{\varphi^{\mathcal{T}}(\lambda\underline{h})}}{4(\log n)^{2+\epsilon}/\varphi^{\mathcal{T}}(\lambda\underline{h})}\right\}^{\gamma}\right) \\ &\quad\le\ L_n\cdot 4\left(1 + \frac{\varepsilon_0^2C\log L_n}{16(\log n)^{2+\epsilon}/\varphi^{\mathcal{T}}(\lambda\underline{h})}\right)^{-\frac{(\log n)^{2+\epsilon}}{2\varphi^{\mathcal{T}}(\lambda\underline{h})}} \\ &\qquad\quad + \frac{16L_nC}{\varepsilon_0}\frac{\sqrt{n\varphi^{\mathcal{T}}(\lambda\underline{h})}}{\log n}\exp\left(-\varsigma\frac{\varepsilon_0^{\gamma}}{4^{\gamma}}\left\{\frac{\sqrt{n\varphi^{\mathcal{T}}(\lambda\underline{h})}}{(\log n)^{1+\epsilon}}\right\}^{\gamma}\right) \\ &\quad\le\ L_n\cdot 4\exp\left(-\frac{\varepsilon_0^2C\log L_n}{32}\right) + CL_n^2\exp(-\varsigma\varepsilon_0^{\gamma}/4^{\gamma}\log L) \\ &\quad\le 4L_n^{-C\varepsilon_0^2/32} + CL_n^{-\frac{\varsigma\varepsilon_0}{4}+2}. \end{aligned} \tag{2.87}$$

Here we used the fact that $\gamma \geq 1$ and (2.43) in Assumption E. Note that in the special case when the response Y_t is assumed to be bounded, the same result continues to hold with $\gamma_1 = \infty$ (so that $\gamma_2 = \gamma(\geq 1)$). Now noting that $\varsigma > 1$, by choosing ε_0 large enough it follows that

$$R_3 = \max_{1\leq k\leq L_n} \left|\widehat{m}_2(x_k) - E\widehat{m}_2(x_k)\right| = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda \underline{h})}}\right). \tag{2.88}$$

Same conclusion holds in the dynamic regression case (i.e. under Assumption A2′) because of the following reason. The penalty term (due to the penalised mixing rate) that incurs inside the curly bracket results in an additional multiplicative term of $\exp(-c(-\tau)) = \exp(c\log n) = n^c$ in the second term of the final bound in (2.87), where $c := \varsigma(\varepsilon_0/4)$ is fixed, and this diverges to infinity at the slower rate than $L_n^{(c-2)} = (n^2/\sqrt{\log n})^{\log n\cdot(c-2)}$.

Returning back to where we started, viewing $\widehat{m}_1(x)$ as a special case of $\widehat{m}_2(x)$ with $Y_t = 1\ \forall t$, we can repeat the above procedure, yielding (since $E\widehat{m}_1(x) = 1$)

$$\sup_{x\in\mathcal{S}_\tau} \left|\widehat{m}_1(x) - 1\right| = O_P\left(\sqrt{\frac{(\log n)^2}{n\varphi(\lambda \underline{h})}}\right). \tag{2.89}$$

The proof is now complete in view of (2.80), (2.81), (2.84), (2.85), (2.88), (2.89), contributions from the bias component, and either Proposition 4.1 of Dunker, Lifshits and Linde (1998) under Assumption C1, or Theorem 1.1 of Hong, Lifshits and Nazarov (2016) under Assumption C2. ■

Proof of Theorem 2.5. Given the extended moment condition upto $8+\delta$, it is straightforward to see (from Theorem 2.1 and 2.2 & 2.3) the consistency of $\widehat{\sigma}^j(x_i)$ for $\sigma^j(x_i)$ for $j = 1, 2, 3, 4$ at every point of continuity x_i, and the asymptotic normality of $(\widehat{\mu}, \widehat{\sigma}^2)$ with limiting variance $\Omega(x_i)$.

Hence it suffices to show asymptotic independence of $\widehat{m}(x_i)$and $\widehat{m}(x_i')$ across i, where x_i and x_i' are continuity points of m such that $\|D^{-1}(x_i - x_i')\| > 0$. Following the notations of the proof of Theorem 2.2 and 2.3, the asymptotic covariance matrix is given by $\mathrm{Var}\big[(\sqrt{\phi(h)}/\sqrt{n})\sum_{t=1}^n R_{nt}\big]$, and

$$\mathrm{Var}(R_{nt}) = \mathrm{Var}\begin{pmatrix} \frac{1}{EK_{1,x}}\cdot K_{t,x}[Y_t - m(x)] \\ \frac{1}{EK_{1,x'}}\cdot K_{t,x'}[Y_t - m(x')] \end{pmatrix} = E\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \tag{2.90}$$

We know from Theorem 2.2 and 2.3 that as for $A_{11} \simeq \sigma^2(x)$ and $A_{22} \simeq \sigma^2(x')$. So we just consider the off-diagonal terms. Due to stationarity we see that

$$
\begin{aligned}
&E\Big[K_{t,x}K_{t,x'}(Y_t - m(x))(Y_t - m(x'))\Big] \\
&\quad = E\Big[K_{1,x}K_{1,x'}\Big[\Big\{(Y_1 - m(X_1)) + (m(X_1) - m(x))\Big\}\Big\{(Y_1 - m(X_1)) + (m(X_1) - m(x'))\Big\}\Big] \\
&\quad = E\Big[K_{1,x}K_{1,x'}(Y_t - m(X_1))(Y_t - m(X_1))\Big] + o(1) = E\Big[K_{1,x}K_{1,x'}\sigma^2(X_1)\Big] + o(1) \\
&\quad \leq \sup_{u \in B(x,h) \cap B(x',h)} \sigma^2(u) E[K_{1,x'}K_{1,x}] \to 0
\end{aligned}
$$

as $h \to 0$ since the kernels return 0 outside its compact support and $\|D^{-1}(x_i - x_i')\| > 0$. The desired result now directly follows via the delta method. ■

Chapter 3

Volatility estimation from high frequency data: A Fourier approach

A financial time series is said to be of *high frequency*, if it is recorded at finer than daily time scale. The extra observations provide us with additional abundant information regarding the price dynamics that may be otherwise not available, and allow us to estimate volatility more accurately. Quoting Zivot (2005), "[The] use of high frequency data has the potential of revolutionizing the way volatility and correlation are modeled and forecasted."

Whereas daily, weekly and monthly observed series have been extensively studied and analyzed in the literature, studies with data sampled at denser frequency have a relatively short history. This is largely due to limited availability of accessible such datasets especially back in and before 1990s. Ever since then however, there has been a remarkable growth in statistical and econometric literature of high frequency data. The development is largely attributed to technological/computational advancements, surging availability of a wide range of data, and rigorous theoretical contributions from the earlier major works. Some examples include Jacod (1994), Foster and Nelson (1996), Jacod and Protter (1998) and Barndorff-Nielsen and Shephard (2002).

High frequency data possess some unique empirical properties that those of lower frequencies do not have. They bring new challenges and distinctive features in the analysis of such time series as we shall study. In this chapter, we consider the estimation problem of the covariance matrices of high frequency asset returns. Our methodology takes into account of the presence of microstructure noise and asynchronicity between the observations across different assets, two major hurdles in high frequency finance. Motivated by Malliavin and Mancino (2002, 2009) we propose a new Fourier domain based estimator of multivariate ex-post volatility, which we call the Fourier Realized Kernel (FRK). An advantage of this approach is that no explicit time alignment is required unlike the time domain based methods widely adopted in the existing literature. We derive the large sample properties and establish asymptotic normality of our estimator under some general conditions that allow for both temporal and cross-sectional correlations in the measurement error process. Our results can be viewed as a Frequency domain extension of the asymptotic theories for the multivariate realized kernel estimator of Barndorff-Nielsen et al. (2011). We show in extensive simulations that our method outperforms the time domain estimators when two assets with different liquidity are traded asynchronously.

3.1 Introduction

Over the past two decades there have been many advances in the theory and application of volatility measurement from high frequency data. The ex-post measure of volatility called the quadratic variation has been the focus of much attention. The theory has been developed in a series of papers including but not limited to: Andersen, Bollerslev, Diebold and Labys (2001), Barndorff-Nielsen and Shephard (2002, 2004), and Mykland and Zhang (2006). This literature has recently been extended to account for what is called the microstructure noise, namely the measurement error that distorts the underlying efficient price diffusion: Zhang, Mykland and Aït-Sahalia (2005), Zhang (2006), Kalnina and Linton (2008), Barndorff-Nielsen, Hansen, Lunde and Shephard (2008, 2011), Aït-Sahalia, Mykland and Zhang (2011), and Bibinger, Hautsch, Malec and Reiss (2014). Jacod, Li, Mykland, Podolskij and Vetter (2009) introduced the pre-averaging method, which involves first averaging the observed prices over a moderate number of time points to reduce the measurement error. In the multivariate case an additional issue arises in the estimation procedure; since transactions most likely occur at different time points for different assets the observations exhibit asynchronicity. Hayashi and Yoshida (2005) proposed an estimator of the integrated covariance that does not require synchronization. However, their estimator is inconsistent under the presence of microstructure noise.

In a seminal paper Malliavin and Mancino (2009) proposed a Fourier domain approach that does not require data alignment, and established consistency and asymptotic mixed normality of their estimator under a general setting (but without deriving the limiting distribution under the presence of measurement error), extending their earlier work Malliavin and Mancino (2002). Since the first version of this work was written, Mancino and Sanfelici (2008) have shown that their univariate estimator is consistent in the presence of measurement error; they also provide mean squared error expansions for their uniform weighting estimator. Furthermore, Curato, Mancino and Sanfelici (2014) derived the bias expression due to microstructure noise, and showed consistency of their estimator. Time domain estimators addressing both asynchronicity and microstructure noise have been proposed by Zhang (2011), Barndorff-Nielsen, Hansen, Lunde and Shephard (2011), and Aït-Sahalia, Fan and Xiu (2010). The estimators are consistent with convergence rates of $O(n^{1/6})$, $O(n^{1/5})$ and $O(n^{1/4})$, respectively. The first two papers require aligning the data, although consistency of their estimator is robust to the alignment. However, the hidden cost of data alignment and non-synchronicity for these estimators is that the sample size n that appears in the convergence rate is the sample size of the aligned data. Also, the drawback of Zhang (2011) and Aït-Sahalia et al. (2010) is that the estimator cannot be generalized to dimensions higher than two unless the covariance matrix is estimated element-wise, which in turn does not guarantee the estimated covariance matrix to be positive definite. See Park and Linton (2012) for a more detailed survey.

The goal of this chapter is to propose an estimator of a general multivariate volatility measure that is robust both to the microstructure noise and asynchronous data timing. The method is based on Fourier domain techniques which have been widely used in discrete time series, and is broadly similar to that of Malliavin and Mancino (2009) although we allow for more general kernel weighting in the Fourier domain. An advantage of this approach is that it does not require an explicit time alignment. The by-product of the Fourier domain based estimator is that we have a consistent estimator of the instantaneous co-volatility even under the presence of quite general dependent microstructure noise. We provide a central limit theorem for our estimator under some general conditions, and also discuss the bandwidth choice issue based on the asymptotic mean squared error expressions. Our results allow for the unbalanced case where one series may have many more observations than another, which is common in practice since stocks vary considerably in terms of their trading intensities. In Section 3.2 we give a setup of the model and assumptions regarding the sampling scheme. In Section 3.3, we propose a Fourier domain based estimator of the integrated covariance. Section 3.4 studies the large sample asymptotics of the proposed estimator and derives its limiting distribution under the presence of microstructure noise. The Fourier method is further extended to estimate the instantaneous covariance matrix of diffusion process. We carry out extensive simulations and empirical analysis, and report the results in Section 3.5.

3.2 Model and assumptions

3.2.1 Efficient Price and Parameter of Interest

The following standard assumption on the efficient price process provides the general framework that will be used throughout this chapter.

ASSUMPTION 3.1. *The efficient price process follows a Brownian semimartingale: For a $p \times 1$ vector of logarithmic prices $P(t) = \big(P_1(t), \ldots, P_p(t)\big)^\intercal$ defined on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, we have*

$$P(t) = \int_0^t \mu(u)du + \int_0^t \sigma(u)dW(u), \tag{3.1}$$

where $\mu(u) = [\mu_1(u), \dots, \mu_p(u)]^\intercal$ is a vector of predictable locally bounded drifts, $\sigma(u)$ is a symmetric $p \times p$ matrix of locally bounded càdlàg process with finite integrated quarticity: $\int_0^t \sigma(u)\sigma(u)^\intercal \otimes \sigma(u)\sigma(u)^\intercal du < \infty$ a.s., and $W(u)$ is a $p \times 1$ vector of Brownian motion that is independent of the volatility process.

The assumptions of locally bounded drift and of diffusion coefficient are needed to apply Girsanov's theorem to remove the drift term in the theoretical derivation. Consider a discrete time grid $0 = t_0 < \cdots < t_n = T$ where T is fixed, and denote by $P(t_i)$ the logarithmic price observed at t_i; $i = 1, ..., n$. The quadratic covariation matrix of P over a time interval $[0, t]$, for some $t \leq T$ is defined by

$$[P, P]_t := \operatorname*{plim}_{n \to \infty} \sum_{i; t_i \leq t} \{P(t_i) - P(t_{i-1})\}\{P(t_i) - P(t_{i-1})\}^\intercal. \tag{3.2}$$

The limit here is finite and well-defined with probability one, and is independent of the choice of the time grid if P is a semimartingale. Furthermore, under Assumption 3.1 one can show that (3.2) is almost surely equal to the integrated covariance matrix:

$$[P, P]_t = \int_0^t \sigma(u)\sigma(u)^\intercal du, \tag{3.3}$$

where $\sigma(u)\sigma(u)^\intercal =: \Sigma(u)$ is the instantaneous covariance matrix. We write $[P, P] := [P, P]_T$, and its j-th diagonal element $[P_j, P_j] = \int_0^T \Sigma_{j,j}(u)du$, the integrated variance of j-th asset. We note that the integrated covariance is related to the covariance matrix of prices by

$$\begin{aligned}
\operatorname{cov}\{P(t)\} &= E\left[\int_0^t \sigma(u)dW(u) \left(\int_0^t \sigma(u)dW(u)\right)^\intercal\right] \\
&= \int_0^t E\left[\sigma(u)\sigma(u)^\intercal\right] du = E[P, P]_t.
\end{aligned}$$

A natural estimator of the quadratic covariation matrix is the realized covariance, the finite sum given in the right hand side of (3.2), which is consistent by construction. Barndorff-Nielsen and Shephard (2002) showed in the univariate framework that the realized variance is an unbiased and $\sqrt{n}-$consistent estimator of the integrated variance under Assumption 3.1.

From a practical viewpoint, two important issues arise in estimating the integrated covariance matrix (3.3). First, in the multivariate setting prices of different assets may be observed at different times, leading to the issue of asynchronicity in observations. Second, observed prices are distorted by some noise due to market microstructure effects, and do not satisfy Assumption 3.1. The objective of this chapter is to propose an estimation theory that is robust to these two problems.

3.2.2 Sampling scheme

In this subsection we describe the main assumptions we make on the observation times. We allow for both *unequal spacing* and *asynchronicity* in random observation times. However, since they are assumed to be strictly exogenous (see Assumption 3.2 below), we shall work with the conditional distributions given the observation times; all statements below should hence be interpreted as stochastic boundedness/convergence.

ASSUMPTION 3.2. *The time span is fixed and is scaled to vary between* $[0, 2\pi]$. *The logarithmic prices are observed at discrete time points:* $0 = t_{0,\ell} < \cdots < t_{n_\ell,\ell} = 2\pi$ *for* $\ell = 1, \ldots, p$, *where* n_ℓ *is the total number of observations for the* ℓ*-th asset. The discrete time points are allowed to be stochastic and are assumed to be independent of the price and volatility process. For asymptotics, we let the smallest number of sample sizes amongst all assets* $n := \min_\ell(n_\ell) \to \infty$. *For* $a, b, \ell \in \{1, \ldots, p\}$:

(1) *The discrete time points satisfy* $\sup_{1\le i\le n_\ell}\big(t_{i,\ell} - t_{i-1,\ell}\big) = s^*/n_\ell = O(n_\ell^{-1})$, *for some finite constant* $s^* > 0$.

(2) *Define the empirical quadratic covariation process of time*: $\forall\ a, b \in \{1, \ldots, p\}$,

$$
\begin{aligned}
\mathcal{Q}^{(n)}_{aabb}(t) &= (n_a \wedge n_b) \sum_{i,j: t_{i,a}, t_{j,b} < t} \Delta t_{i,a} \Delta t_{j,b} 1_{\{I_{i,a}\cap I_{j,b}\neq\emptyset\}} \\
\mathcal{Q}^{(n)}_{abab}(t) &= (n_a \wedge n_b) \sum_{i,j,\ell: t_{i,a}, t_{j,b}, t_{l,b} < t} (t_{i,a} \wedge t_{j,b} - t_{i-1,a} \vee t_{j-1,b}) \\
&\qquad \times (t_{i,a} \wedge t_{l,b} - t_{i-1,a} \vee t_{l-1,b}) 1_{\{I_{i,a}\cap I_{j,b}\neq\emptyset\}} 1_{\{I_{i,a}\cap I_{\ell,b}\neq\emptyset\}}, \ \textit{for } n_a < n_b.
\end{aligned}
$$

The empirical quadratic covariation satisfies $\mathcal{Q}^{(n)}_{abab}(t) \longrightarrow \mathcal{Q}_{abab}(t)$ as $n_a \wedge n_b \to \infty$, *where the limit* $\mathcal{Q}_{abab}(t)$ *is continuously differentiable in* t. *Similarly for* $\mathcal{Q}_{abab}(t)$.

REMARK. Denote the intervals $I_{i,a} = [t_{i-1,a}, t_{i,a})$ and $I_{j,b} := [t_{j-1,b}, t_{j,b})$. Assumption 3.2-(1) implies that first, the degree of non-synchronicity satisfies

$$
\sup_{i,j} |t_{i,a} - t_{j,b}| 1_{\{I_{i,a}\cap I_{j,b}\neq\emptyset\}} = O\left(\frac{1}{n_a \wedge n_b}\right).
$$

Further, given any set of bivariate time points $\{t_{i,a}, t_{j,b}\}$ with $n_a < n_b$, we have

$$
\sup_{0\le j\le n_b} \#\left\{t_{j,b} \in [t_{i-1,a}, t_{i,a}) | 1_{\{I_{i,a}\cap I_{j,b}\neq\emptyset\}}\right\} = O\left(\frac{n_a \vee n_b}{n_a \wedge n_b}\right).
$$

We note that the arrays $\mathcal{Q}_{abab}(t)$ and $\mathcal{Q}_{aabb}(t)$ will appear in the asymptotic variance of our estimator for the integrated variance. In the univariate case, the expression in

Assumption 3.2-(2) specializes to $\mathcal{Q}_{aa}^{(n)}(t) = n_a \sum_{i:t_{i,a}<t}(\Delta t_{i,a})^2$. It is also worth noting that $\{1_{\{I_{i,a}\cap I_{j,b}\neq\emptyset\}} = 1\}$ if and only if $\{u_{ij} := (t_{i,a} \wedge t_{j,b}) > (t_{i-1,a} \vee t_{j-1,b}) =: l_{i,j}\}$. Lastly, we see that the assumption does not restrict the ratio of sample sizes of different assets to be bounded away from zero or infinity (see also Assumption 3.4). In summary, both unequal spacing and asynchronicity in observation are allowed in a sufficiently general way.

We introduce some notations that will be used in the sequel. Define

$$\big\{T_l^{(ab)}\big\}_{1\leq l\leq N_T^{(ab)}} := \big\{t_{i,a} \cup t_{j,b}; i = 1,\ldots,n_a, j = 1,\ldots,n_b\big\}, \tag{3.4}$$

where $N_T^{(ab)}(\leq n_a + n_b)$ is the total number of data points for the union of time stamps. Denote the average interval size for asset ℓ by $\overline{\Delta t}_\ell := 2\pi/n_\ell$. As for two different assets say a and b that are being compared, the average interval size of the more liquid asset is denoted by $\widetilde{\Delta t}_{ab} := 2\pi/(n_a \vee n_b)$.

3.3 Estimation

3.3.1 The Fourier Kernel Estimator

Motivated by the disadvantages of data alignment methods widely discussed in the literature, see for example Aït-Sahalia and Jacod (2014) for relevant discussions, we propose to use a Fourier domain approach which does not require data synchronization. Our framework follows the line of approach of Malliavin and Mancino (2002, 2009), where a nonparametric method based on Fourier analysis of returns was discussed. Frequency domain techniques are widely used in discrete time series analysis; an important application of such an approach is the estimation problem of the long-run variance of a stationary time series (which is proportional to the spectral density at zero frequency). We draw a natural link between those classical theories and the estimation of the quadratic covariation of a continuous time processes.

The Fourier basis $\{g_t(q) := e^{\mathrm{i}qt}/\sqrt{2\pi}; q \in \mathbb{Z}\}$ where $\mathrm{i} = \sqrt{-1}$ constitutes an orthonormal basis of $L^2([0, 2\pi])$:

$$\frac{1}{2\pi}\int_0^{2\pi} g_t(k)\overline{g_t(j)}dt = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{otherwise,} \end{cases}$$

where $\overline{g_t(q)}$ refers to the complex conjugate of $g_t(q)$.

In view of completeness of the Fourier basis, we can express the continuous time signal $\Sigma(t) \in L^2([0, 2\pi])$ as a linear combination of the Fourier basis with coefficients denoted by $\mathcal{F}(\Sigma)(q)$:

$$\Sigma(t) = \frac{1}{2\pi} \sum_{q=-\infty}^{\infty} \mathcal{F}(\Sigma)(q) e^{\mathrm{i}qt}, \tag{3.5}$$

where

$$\mathcal{F}(\Sigma)(q) = \int_0^{2\pi} e^{-\mathrm{i}qt} \Sigma(t) dt; \quad q \in \mathbb{Z}. \tag{3.6}$$

This is the continuous time Fourier transform of the instantaneous covariance matrix. Note that (3.6) corresponds to the integrated covariance when $q = 0$; we will propose an estimator for the general q case. The Fourier pair above suggests that once we estimate the Fourier coefficient by a suitable estimator $\widehat{\mathcal{F}}(\Sigma)(q)$, the signal can be reconstructed via:

$$\widehat{\Sigma}(t) = \frac{1}{2\pi} \sum_{q=-n}^{n} \widehat{\mathcal{F}}(\Sigma)(q) e^{\mathrm{i}qt}.$$

We now discuss the estimation procedure of (3.6) based on Fourier transform of the return process. Define

$$\mathcal{F}(dP_\ell)(\alpha) = \int_0^{2\pi} e^{-\mathrm{i}\alpha t} dP_\ell(t); \quad \alpha \in \mathbb{Z},\ \ell = 1, ..., p, \tag{3.7}$$

where $P_\ell(t)$ refers to observed price of the ℓ-th asset satisfying Assumption 1, and let

$$\mathcal{F}_n(dP_\ell)(\alpha) = \sum_{j=1}^{n_\ell} e^{-\mathrm{i}\alpha t_{j,\ell}} \Delta P_\ell(t_{j,\ell}), \tag{3.8}$$

whose vector version $\mathcal{F}_n(dP)(\alpha) = \left\{\mathcal{F}_n(dP_1)(\alpha), \ldots, \mathcal{F}_n(dP_p)(\alpha)\right\}^\intercal$, $\alpha \in \mathbb{Z}$, can be defined to account for multiple assets altogether.

Now we consider a class of kernel called the spectral window $K_H : [-\pi/H, \pi/H] \to \mathbb{R}$, where $K_H(\lambda) \simeq H \cdot K(H\lambda)$ and $K(\cdot)$ is a function called the spectral window generator. Detailed analysis of these weighting functions is given in the next subsection (Assumption 3.3′, (3.14) and relevant discussions thereof); see also Parzen (1967) and Subba Rao and Gabr (1984) for earlier discussions. The estimator we propose for the integrated covariance (3.6) is given by

$$\widehat{\mathcal{F}}(\Sigma)(q) = \left(\widehat{\mathcal{F}}(\Sigma_{i,j})(q)\right) = \sum_{|\alpha| \le m/2} K_H(\lambda_\alpha) \Big[\mathcal{F}_n(dP)(\alpha)\Big] \Big[\mathcal{F}_n(dP)(q-\alpha)\Big]^\intercal, \tag{3.9}$$

where $\lambda_\alpha = 2\pi\alpha/\overline{n}$ and $\overline{n} := \max_{\ell=1,...,p} n_\ell$. Let $m = \overline{n}/H$ where the bandwidth $H = H_n \to \infty$ but not as fast as $n = \min_\ell n_\ell \to \infty$ so that $m \to \infty$. We are smoothing over the interval $[-\pi/H, \pi/H]$, where H controls the width of the smoothing window. We shall refer to our estimator as the *Fourier Realized Kernel* (FRK). Note that the estimator considered in Malliavin and Mancino (2009) can be compared with ours with

$K(x) = (1/\pi)(1 - |x|/\pi)$.

The $q = 0$ case deserves a special attention. In this case $\mathcal{F}_n(dP_1)(\alpha)\mathcal{F}_n(dP_2)(-\alpha) =: I_{12}(\alpha)$ is the realized cross periodogram between assets 1 and 2, say. Then, the $(1,2)$-th entry of $\widehat{\mathcal{F}}(\Sigma)(0)$ is given by kernel smoothing the realized cross periodogram around zero frequency:

$$\widehat{\mathcal{F}}(\Sigma_{12})(0) = \sum_{|\alpha| \leq m/2} K_H(\lambda_\alpha) I_{12}(\alpha). \tag{3.10}$$

Positive definiteness of the estimators (3.10) is guaranteed provided that the spectral window is non-negative.

3.3.2 Comparison with some Time domain estimators

For data that is synchronized at $\{\tau_i\}$, its realized autocovariance function is defined as

$$\gamma_{12}(h) = \sum_i \Delta P_1(\tau_i)\Delta P_2(\tau_{i-h}); \quad h \in \mathbb{Z}, \tag{3.11}$$

where $\sum_i = \sum_{h<i\leq n}$ for $h \geq 0$, and $\sum_i = \sum_{1 \leq i \leq n+h}$ for $h < 0$. In the aligned case the realized periodogram is closely related to the realized autocovariance. In particular when τ_i are equally spaced and synchronous, i.e. $\tau_i = \tau_j + (i-j)2\pi/n$, it can be easily shown that the realized cross periodogram is the Fourier transform of the realized autocovariance; that is, $I_{12}(\alpha) = \sum_{|h|<n} e^{-\mathrm{i}\alpha h 2\pi/n}\gamma_{12}(h)$.

Hayashi and Yoshida (2005) considered a covariation estimator defined as the realized cross periodogram at zero frequency over the interval that overlaps:

$$HY = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \Delta P_1(t_{i,1})\Delta P_2(t_{j,2}) 1_{\{I_{i,1} \cap J_{j,2} \neq \emptyset\}}.$$

Then the estimator can be expressed in terms of the centered (i.e. frequency zero) realized cross periodogram $I_{1,2}(0)$. In particular, we have the following decomposition:

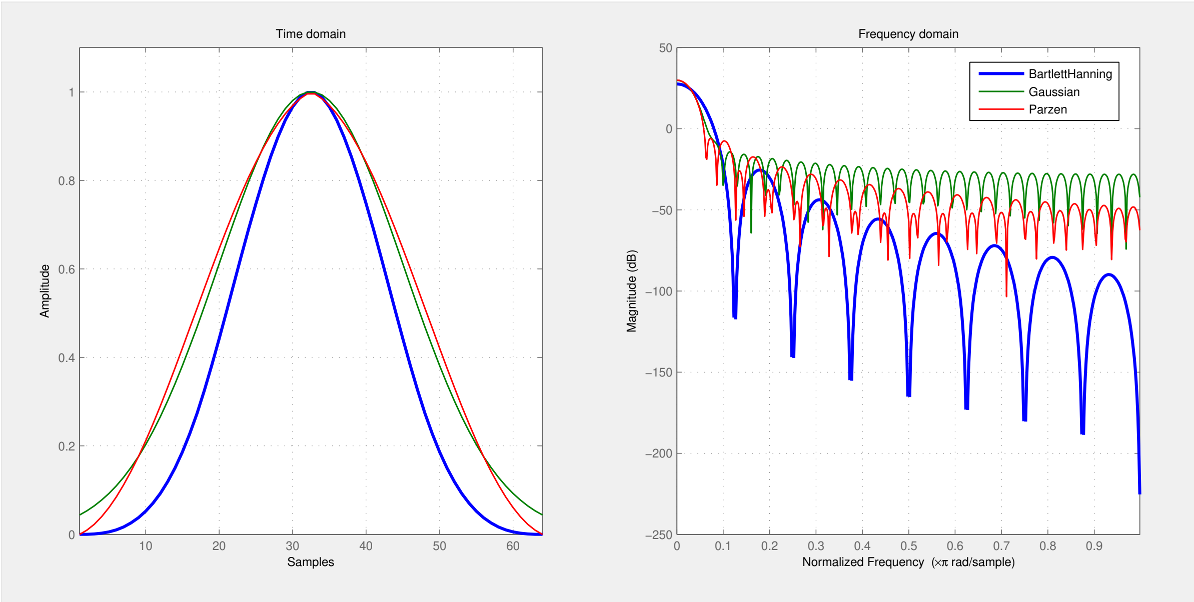
$$I_{1,2}(0) - [P_1, P_2] = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \Delta P_1(t_{i,1})\Delta P_2(t_{j,2}) - \int_0^{2\pi} \Sigma_{12}(t)dt = M_1 + M_2,$$

where

$$M_1 = HY - \int_0^{2\pi} \Sigma_{12}(t)dt, \quad M_2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \Delta P_1(t_i)\Delta P_2(s_j) 1_{\{I_{i,1} \cap J_{j,2} = \emptyset\}}.$$

Hayashi and Yoshida (2008) showed that $\sqrt{n}M_1$ is asymptotically zero mean Gaussian (when data is Poisson sampled), and M_2 is mean zero with leading term of order $O_p(1)$. Hence, provided there is no microstructure noise the Hayashi and Yoshida estimator is

Figure 3.1: Examples of the lag and spectral windows satisfying Assumption 3 and 3′



unbiased, and achieves $\sqrt{n}$-consistency without requiring synchronization. The realized periodogram is also unbiased but is inconsistent due to the presence of M_2.

We next compare our estimator (3.10) with the (multivariate) *realized kernel* estimator by Barndorff-Nielsen et al. (2008, 2011), denoted by $\widetilde{\Sigma}$. It is given by kernel smoothing the realized autocovariances of the data aligned on $\{\tau_i\}_{i=1}^N$ using the refresh time sampling scheme. For example, its $(1,2)$th entry is

$$\widetilde{\Sigma}_{12} := \sum_{|h|<n} k\left(\frac{h}{H}\right)\gamma_{12}(h) = \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta P_1(\tau_i)\Delta P_2(\tau_j) k\left(\frac{i-j}{H}\right), \tag{3.12}$$

where $k(\cdot)$ is a smoothing window in time domain. To establish the link between the time domain estimator (3.12) and our frequency domain estimator (3.10), we discuss the properties of smoothing windows; we introduce the lag window for continuous time denoted by $k(x)$, $x \in \mathbb{R}$ and compare it with the spectral kernel for continuous and bandlimited frequency. The lag window k is assumed to satisfy the following conditions (introduced as Assumption K in Barndorff-Nielsen et al. (2011)). Note that the prime notation is taken to mean differentiation with respect to the argument.

Assumption 3.3. *The lag window $k(\cdot)$ satisfies the following conditions: (i) k is twice continuously differentiable; (ii) $k(0) = 1$, and $k'(0) = 0$; (iii) $\|k\|^2 := \int_{-\infty}^{\infty} |k(x)|^2 dx < \infty$, $\|k^2\|^2 := \int_{-\infty}^{\infty} |k(x)|^4 dx < \infty$, $\|k'\|^2 := \int_{-\infty}^{\infty} |k'(x)|^2 dx < \infty$, $\|k''\|^2 := \int_{-\infty}^{\infty} |k''(x)|^2 dx < \infty$; (iv) $\int_{-\infty}^{\infty} k(x)\exp(-\mathrm{i}\lambda x)dx \geq 0$, $\forall \lambda \in [-\pi, \pi]$.*

We define the spectral window generator K as a Fourier transform of the lag window k:

$$K(\lambda) = \tfrac{1}{2\pi}\int_{-\infty}^{\infty} k(t)e^{-\mathrm{i}\lambda t}dt \qquad k(t) = \int_{-\pi}^{\pi} K(\lambda)e^{\mathrm{i}\lambda t}d\lambda, \tag{3.13}$$

where λ denotes the angular frequency. Then it follows from simple algebra and Parseval's identity that Assumption 3.3 on the lag window k translates to the following conditions on the spectral window generator.

Assumption 3.3′. *The spectral window generator $K(\cdot)$ satisfies the following conditions:* (*i*) $\int_{-\pi}^{\pi} K(\lambda)d\lambda = 1$, $\int_{-\pi}^{\pi} \lambda K(\lambda)d\lambda = 0$; (*ii*) $\|K\|^2 := \int_{-\pi}^{\pi} |K(\lambda)|^2 d\lambda < \infty$, $\mu_1^2(K) := \int_{-\pi}^{\pi} |\lambda K(\lambda)|^2 d\lambda < \infty$ *and* $\mu_2^2(K) := \int_{-\pi}^{\pi} |\lambda^2 K(\lambda)|^2 d\lambda < \infty$; (*iii*) $K(\lambda) \geq 0$, $\forall\lambda \in [-\pi, \pi]$.

Figure 3.1 shows some examples of weighting functions k and K that satisfy Assumption 3.3 and 3.3′. In the sequel, Assumption 3.3 is taken to mean *both* Assumption 3.3 *and* Assumption 3.3′ collectively.

As remarked by Barndorff-Nielsen et al. (2011), condition (*iv*) in Assumption 3.3 (and equivalently (*iii*) in Assumption 3.3′) guarantees positive semi-definiteness of the estimators defined in (3.12) and (3.10) via Bochner's theorem. The realized periodogram is also positive semi-definite and is Hermitian as long as the spectral window is non-negative.

Consider the discrete time and discrete frequency Fourier pair:

$$K_H(\lambda_\alpha) = \tfrac{1}{2\pi}\textstyle\sum_{|h|\leq H} k\left(\tfrac{h}{H}\right) e^{\mathrm{i}\lambda_\alpha h}; \qquad k\left(\tfrac{h}{H}\right) = \textstyle\sum_{\alpha=-m/2}^{m/2-1} K_H(\lambda_\alpha)e^{-\mathrm{i}\lambda_\alpha h}. \tag{3.14}$$

We refer to the weighting function K_H defined in (3.14) as the spectral window. To avoid the aliasing problem we assume that the signal is zero for frequencies that falls outside of the Nyquist critical frequency. As shown in Parzen (1967, page 130), the spectral window that is used to smooth the realized periodogram is related to a spectral window generator via:

$$K_H(\lambda) = H\sum_{j=-\infty}^{\infty} K\big(H(\lambda + 2\pi j)\big) \simeq HK(H\lambda).$$

Now with $\lambda_\alpha = 2\pi\alpha/(n_a \vee n_b)$, $\alpha \in \mathbb{Z}$ for some $a, b = 1, ..., p$, upon substituting $h = (n_a \vee n_b) \cdot (t_i - s_j)/(2\pi)$ into the latter expression of (3.14) we see that

$$\begin{aligned}\sum_{|\alpha|<m/2} K_H(\lambda_\alpha)I_{ab}(\alpha) &= \sum_{i=1}^{n_a}\sum_{j=1}^{n_b} \Delta P_a(t_i)\Delta P_b(s_j) \sum_{|\alpha|<m/2} K_H(\lambda_\alpha)e^{-\mathrm{i}\alpha(t_i-s_j)} \\ &= \sum_{i=1}^{n_a}\sum_{j=1}^{n_b} \Delta P_a(t_i)\Delta P_b(s_j)k_H(t_i - s_j), \end{aligned} \tag{3.15}$$

where $k_H(t_i - s_j) := k((t_i - s_j)/(\widetilde{\Delta t}_{ab} H))$, where $\widetilde{\Delta t}_{ab} := 2\pi/(n_a \vee n_b)$. Further, since an equally spaced and synchronized time grid satisfies $t_i = s_j + (i-j)2\pi/(n_a \vee n_b)$, the following statement of key importance trivially holds:

REMARK 3.1. *When trading times are synchronized and equally spaced, the Fourier realized kernel at zero frequency* (3.10) *and the multivariate realized kernel* (3.12) *are identical.*

We will show later that when the data is not synchronously observed, using all the data and implementing (3.15) delivers a superior estimator.

It is of interest how our Fourier kernel estimator is related to other time domain estimators such as the multivariate two time scale estimator of Zhang (2011), and the modulated realized covariance (multivariate pre-averaging estimator) of Christensen, Kinnebrock and Podolskij (2010). In the univariate setting, Jacod et al. (2009) showed that their pre-averaging estimator, the univariate two time scale estimator of Zhang et al. (2005), and the flat-top realized kernel of Barndorff-Nielsen et al. (2008) can be written as a smoothed realized autocovariances, and the difference between the estimators comes from the contribution of the end points. This result holds also for the multivariate versions of the three estimators when observation points are synchronized. Our estimator can be expressed as a realized kernel only when sampling points are equally spaced and aligned. The relation between the smoothed periodogram to estimate the spectrum and data tapering (i.e. Fourier transforming the weighted return) is analogous to the relation between our estimator and the pre-averaging estimator.

3.4 Asymptotic Properties

3.4.1 Bandwidth conditions

We introduce the rate conditions on the bandwidth we require for our asymptotic theories. Our conditions allow the sample sizes of different assets to be of different order of magnitude. As aforementioned, such a situation arises often in practice as some assets may be traded much more frequently than others. As defined previously in Assumption 3.2, n is taken to mean the minimum amongst the individual sample sizes of the assets throughout, unless stated otherwise.

ASSUMPTION 3.4. *The bandwidth sequence $H = H_n$ is of order $H = O(n^\kappa)$ where $\kappa \in (0,1)$ so that $H \to \infty$ and n/H $(=: m) \to \infty$ as $n \to \infty$. Moreover, the ratio of the bigger to smaller sample size between two assets satisfies $(n_a \vee n_b)/(n_a \wedge n_b) = o(H)$ for any $a, b \in \{1, \ldots, p\}$.*

The first condition regulates the speed at which the bandwidth is allowed to increase; this is consistent with the usual setting in nonparametric literature. The next requirement $(n_a \vee n_b)/(H(n_a \wedge n_b)) = o(1)$ specifies the asymptotic behaviour of the ratio of the larger to smaller sample sizes; this will be shown to control the bias from the leading term. We define β $(= \max_{a,b} \beta_{ab}) \geq 1$ to be the degree of liquidity parameter where

$$\beta_{ab} = \lim_{n \to \infty} \frac{\log (n_a \vee n_b)}{\log (n_a \wedge n_b)}. \tag{3.16}$$

The definition suggests that if we write $n_a \wedge n_b =: n_{ab}$, then we have $n_a \vee n_b = n_{ab}^{\beta_{ab}}$; this in turn implies that $1 \leq \beta_{ab} < 2$ for all $a, b \in \{1, \dots, p\}$ because otherwise $(n_a \vee n_b)/(H(n_a \wedge n_b))$ would not converge, violating Assumption 3.4. So essentially, Assumption 3.4 implies that given $n_a > n_b$, we have $n_a << n_b^2$. As a slight abuse of notation, we will hereafter omit the subscript of β whenever it is clear from the context which assets are being considered.

3.4.2 Assumption on the microstructure noise

Empirical evidence from the volatility signature plot suggests that the observed price deviates from the semimartingale assumption. In fact, various studies document that the observed high frequency returns have infinite quadratic variation. The following assumption is proposed to account for this phenomena.

Assumption 3.5. *The observed logarithmic price of the ℓ-th asset, denoted $X_\ell(t_{j,\ell})$, is the sum of two components. The first component is a discretely observed continuous signal $P_\ell(t_{j,\ell})$ that satisfies Assumption 1, and the other one is a noise process with respect to the realization of transaction time $U_\ell(t_{j,\ell})$ that has an infinite quadratic variation: i.e. for each $\ell = 1, ..., p$*

$$X_\ell(t_{j,\ell}) = P_\ell(t_{j,\ell}) + U_l(t_{j,\ell}). \tag{3.17}$$

This type of additive microstructure noise model has been well-studied in the literature, particularly in the univariate context; see for recent examples, Mancini and Sanfelici (2008) and Curato et al. (2014) where the noise $U_\ell(t_{j,\ell})$ is assumed to be i.i.d. There has not been much empirical work that studied cross autocorrelation of the microstructure noise for multiple assets. Amongst the few includes Aït-Sahalia et al. (2010) where the noise is set to be i.i.d in time but uncorrelated across different assets. Zhang (2011) assumed (covariance-) stationarity and an exponential alpha mixing condition with respect to the observation time.

In this chapter, we shall impose the following mild dependence condition on the microstructure noise:

Assumption 3.6. *The stochastic process $U_\ell(.)$ is stationary, mean zero, and is independent of the efficient price process $P_\ell(.)$. Furthermore, the covariance function of the noise process defined as $E[U_a(t_{i,a})U_b(t_{j,b})] =: \gamma(|t_{i,a} - t_{j,b}|/\widetilde{\Delta t}_{ab})$ satisfies*

$$\frac{1}{n_a \wedge n_b} \sum_{i=1}^{n_a-1} \sum_{j=1}^{n_b-1} \gamma\left(\frac{|t_{i,a} - t_{j,b}|}{\widetilde{\Delta t}_{ab}}\right) \longrightarrow \Gamma_{ab} \ (< \infty), \tag{3.18}$$

where the finite limit Γ_{ab} is the (a,b)-th element of the $p \times p$ positive semi-definite covariance matrix Γ. Also, the fourth moments satisfy $|\mathrm{E}(U_a(t_{i,a})U_b(t_{j,b})U_c(t_{r,c})U_d(t_{l,d}))| \leq \rho(M) < \infty$, where $M := \sup_{u,v,g,s}[(t_{u,g} - t_{v,s})/\widetilde{\Delta t}_{gs}]$ and $\rho(\cdot)$ is a function such that for some $\delta > 0$,

$$\sum_{\nu=1}^{\infty} \rho(\nu)(1+\delta)^{\nu} < \infty. \tag{3.19}$$

Our assumption allows both cross-sectional and temporal correlations in the measurement error process. Note that when the data is equally spaced and balanced, (3.18) simplifies to

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma(i-j) \to \Gamma_{ab} < \infty,$$

a condition implied by the standard absolute summability condition for the autocovariance function, which is well-known to be sufficient for ergodicity and necessary for (a certain class of) mixing under stationarity, see Ibragimov and Linnik (1971). The second condition (3.19) implies that the supremum of the fourth moment decays sufficiently fast as the maximum cross-lag $M = \max\big(|i-j|, |h-l|\big)$ increases.

3.4.3 Asymptotic Normality

We now introduce our main theoretical results. At the end, the limiting distribution of our estimator (3.9) will be derived under the presence of the microstructure noise. Before we proceed, we impose a condition on the end points. This condition is in line with what is assumed in the existing literature in order to ensure consistency of the estimator at the end points, see Jacod et al. (2009) and Barndorff-Nielsen et al. (2011, Section 2.2).

Assumption 3.7. *The observed prices at two end points, i.e. $X_\ell(t_{0,\ell})$ and $X_\ell(t_{n_\ell,\ell})$ for each $\ell = 1, ..., p$ are respectively the arithmetic average of m number of distinct observations on the interval $[t_{-1,\ell}, t_{0,\ell})$ and $[t_{n_\ell,\ell}, t_{n_\ell+1,\ell})$. The time points $t_{-1,\ell}$ and $t_{n_\ell+1,\ell}$ satisfy*

Assumption 3.2-(1) and $m = o(n_\ell)$.

Since Assumption 3.5 implies $\Delta X(t_i) = \Delta P(t_i) + \Delta U(t_i)$, our estimator (3.9) can be decomposed as follows:

$$\begin{aligned}
&\widehat{\mathcal{F}}(\Sigma)(q) - \mathcal{F}(\Sigma)(q) \\
&\quad= \sum_{|\alpha|\le m/2} K_H(\lambda_\alpha)\Big\{\Big[\mathcal{F}_n(dP)(\alpha)\Big]\Big[\mathcal{F}_n(dP)(q-\alpha)\Big]^\intercal - \Big[\mathcal{F}(dP)(\alpha)\Big]\Big[\mathcal{F}(dP)(q-\alpha)\Big]^\intercal\Big\} \\
&\qquad+ \sum_{|\alpha|\le m/2} K_H(\lambda_\alpha)\Big[\mathcal{F}(dP)(\alpha)\Big]\Big[\mathcal{F}(dP)(q-\alpha)\Big]^\intercal - \mathcal{F}(\Sigma)(q) \\
&\qquad+ \sum_{|\alpha|\le m/2} K_H(\lambda_\alpha)\Big[\mathcal{F}_n(dU)(\alpha)\Big]\Big[\mathcal{F}_n(dU)(q-\alpha)\Big]^\intercal \\
&\qquad+ \sum_{|\alpha|\le m/2} K_H(\lambda_\alpha)\Big\{\Big[\mathcal{F}_n(dU)(\alpha)\Big]\Big[\mathcal{F}_n(dP)(q-\alpha)\Big]^\intercal + \Big[\mathcal{F}_n(dP)(\alpha)\Big]\Big[\mathcal{F}_n(dU)(q-\alpha)\Big]^\intercal\Big\} \\
&\quad= (i) + (ii) + (iii) + (iv). \qquad (3.20)
\end{aligned}$$

The first term (i) is the discretization error due to sampling the continuous time signal at discrete points, and (ii) is the error due to kernel smoothing. Moreover, (iii) can be interpreted as the contribution from the smoothed realized periodogram applied to a microstructure noise. Lastly, (iv) is the sum of cross terms between the efficient price and the noise, which is of a smaller order than (iii), see Theorem 1 of Barndorff-Nielsen et al. (2011). Below we derive the asymptotic order of the noise contribution (iii), and establish asymptotic normality of $(i)+(ii)$. All proofs are contained in the appendix.

Proposition 3.1. *Suppose Assumptions 3.2-3.7 hold. Then for any* $a, b = 1, ..., p$,

$$E\left\{\sum_{|\alpha|\le m/2} K_H(\lambda_\alpha)\Big[\mathcal{F}_n(dU_a)(\alpha)\Big]\Big[\mathcal{F}_n(dU_b)(q-\alpha)\Big]\right\} \simeq \left(\frac{s^*}{2\pi}\right)^2 \frac{n_a \vee n_b}{H^2}|k''(0)|\Gamma_{ab}$$

$$E\left\{\left|\sum_{|\alpha|\le m/2} K_H(\lambda_\alpha)\Big[\mathcal{F}_n(dU_a)(\alpha)\Big]\Big[\mathcal{F}_n(dU_b)(q-\alpha)\Big]\right|^2\right\} = O\left(\frac{(n_a \vee n_b)^2}{n_a \wedge n_b} H^{2\mu-3}\right),$$

where μ *is some constant such that* $0 < \mu < 1$, *and* s^* *and* Γ_{ab} *are as defined in Assumption 3.2 and Assumption 3.6, respectively.*

The proposition above implies that the leading bias term (iii) is of order $O_p((n_a \vee n_b)/H^2)$; a necessary condition for this bias contribution from the microstructure noise to be negligible in large sample is $\kappa \in (1/2, 1)$.

We now move on to the main term $(i)+(ii) =: \widehat{\mathcal{F}}(\Sigma^P)(q) - \mathcal{F}(\Sigma)(q)$, and derive its asymptotic bias and variance expressions.

Proposition 3.2. *Suppose Assumptions 3.1-3.4 hold. Then the bias and variance of the estimator $\big(\widehat{\mathcal{F}}(\Sigma^P)(q)\big)_{ab}$ $(a, b = 1, ..., p)$ denoted respectively by $\boldsymbol{B}_{ab}$ and $\boldsymbol{V}_{ab}$ are given by:*

$$\begin{aligned}
\boldsymbol{B}_{ab} &= \left(\frac{n_a \vee n_b}{(n_a \wedge n_b)H}\right)^2 \frac{1}{2}\mathcal{A}_{ab}^2 \left|k''(0)\right| \int_0^{2\pi} e^{-itq} d\left|[P_a, P_b]\right|(t) \\
\boldsymbol{V}_{ab} &= \frac{H}{n_a \wedge n_b}\|k\|^2 \left[\int_0^{2\pi} e^{-i2tq}[P_a, P_a]'(t)[P_b, P_b]'(t) d\mathcal{Q}_{aabb}(t)\right. \\
&\qquad\qquad \left. + \int_0^{2\pi} e^{-i2tq}([P_a, P_b]')^2(t) d\mathcal{Q}_{abab}(t)\right],
\end{aligned}$$

with $\boldsymbol{B}_{aa} = 0$ (for all $a = 1, ..., p$), where

$$\mathcal{A}_{ab} := \lim_{n_a \wedge n_b \to \infty} \frac{n_a \wedge n_b}{2\pi} \sup_{i,j} |t_i - s_j| 1_{\{I_{i,a} \cap I_{j,b} = \emptyset\}}.$$

This result gives the asymptotic mean squared error of our estimator in the absence of microstructure noise but with asnychronous trading. It may be compared with Barndorff-Nielsen et al. (2011) Theorem A.5. Malliavin and Mancino (2009) Theorem 4.1 gives a CLT for their estimator in the absence of microstructure noise. Our result in Proposition 3.2 holds under different restrictions on the bandwidth sequence, and so is not directly comparable.[1]

We note that $\mathcal{A}_{ab}$ can be thought of as a measure of the degree of non-synchronicity between two assets a and b; Assumption 3.2 implies that $\mathcal{A}_{ab} \in [0, \infty)$. In particular, when two series in consideration are perfectly synchronized and balanced, then $\mathcal{A}_{ab} = 0$. Otherwise it is a bounded constant i.e. $O(1)$ by Assumption 3.2.

From Proposition 3.2 we see that the order of asymptotic variance is $O(H/(n_a \wedge n_b))$. The squared bias is given by $O((n_a \vee n_b)^2/H^4)$ as this dominates that of the main term. The optimal bandwidth $H := O(n^\kappa)$ is obtained by balancing those two quantities. Denote by $n = (n_a \wedge n_b) = n_a$ say, and let $n_b = (n_a \vee n_b) = O(n^\beta)$. Then solving $n^{2\beta - 4\kappa} = n^{\kappa - 1}$ we have

$$H = C_0 n^{\kappa_{\text{opt}}}, \quad \kappa_{\text{opt}} = \frac{2\beta + 1}{5}, \tag{3.21}$$

where C_0 is some positive constant. In particular, when two sample sizes n_a and n_b are of the same order (i.e. $\beta = 1$), then the optimal order of the bandwidth is given by $\kappa_{\text{opt}} = 3/5$.

[1] Their estimator has a specific weighting and models asynchronicity in a slightly different way.

The rate of convergence of the estimator is therefore $(n_a \wedge n_b)^{\vartheta}$, where $\vartheta = (2-\beta)/5 \in (0, 1/5]$. This suggests that the best rate $(n_a \wedge n_b)^{1/5}$ is achieved when the sample sizes are of the same order, and the rate slows down as $\beta \in [1,2)$, the degree of relative liquidity between assets in consideration increases.

Theorem 3.1. *Suppose that Assumptions 3.1-3.7 hold. Then provided that the optimal bandwidth (3.21) is chosen, we have for each $q \in \mathbb{Z}$*

$$\mathcal{D}_n \mathrm{vech}\Big\{\widehat{\mathcal{F}}(\Sigma)(q) - \mathcal{F}(\Sigma)(q)\Big\} \Longrightarrow^{stably} N\big(\mathcal{B}, \mathcal{V}\big)$$

as $n \to \infty$, where

$$\begin{aligned}
\mathcal{D}_n =& \mathrm{diag}\left\{\mathrm{vech}(\mathcal{D}_n^*)\right\}, \quad \{\mathcal{D}_n^*\}_{a,b} = (n_a \wedge n_b)^{\vartheta}, \quad \vartheta = \frac{2-\beta}{5}, \\
\mathcal{B} =& (2\pi)^{-2} C_0^{-2} s^{*,2} |k''(0)| \cdot \mathrm{vech}\big(\Gamma\big), \qquad \mathcal{V} = (\mathcal{V}_{ab}), \\
\mathcal{V}_{ab} =& \|k\|^2 \int_0^{2\pi} e^{-i2tq} \Big\{[P_a, P_a]'(t)[P_b, P_b]'(t) d\mathcal{Q}_{aabb}(t) + ([P_a, P_b]')^2(t) d\mathcal{Q}_{abab}(t)\Big\}.
\end{aligned}$$

3.4.4 Estimation of the Instantaneous covariance matrix

Here we extend the Fourier method discussed above to estimate the instantaneous covariance matrix of the diffusion process under the presence of microstructure noise. We can construct an estimator of the instantaneous covariation matrix by Fourier inverting (3.9):

$$\widehat{\Sigma}(t) = \frac{1}{2\pi} \sum_{|q| \le m/2} K_H(\lambda_q) \exp(\mathrm{i}qt) \widehat{\mathcal{F}}(\Sigma)(q). \tag{3.22}$$

Suppose that the modulus of continuity of $\Sigma(t)$ denoted by $\mathcal{C}(h)$ satisfies

$$\mathcal{C}(h) := \sup_{|t-s| \le h} \|\Sigma(t) - \Sigma(s)\|_2 \longrightarrow 0 \tag{3.23}$$

as $h \to 0$. The continuity assumption is met when each element of $\Sigma(t)$ in Assumption 3.1 does not contain jumps, for example $\Sigma(t)$ is a Brownian semimartingale.

Theorem 3.2. *Suppose that the regularity conditions of Theorem 3.1 and (3.23) hold. Then, there exists a sequence $\delta(n) \to 0$ such that*

$$\lim_{n \to \infty} \sup_{\delta(n) \le t \le 2\pi - \delta(n)} \|\widehat{\Sigma}(t) - \Sigma(t)\|_2 = 0.$$

Practitioners often encounter a problem of running a regression between variables that are asynchronously observed; for example, we might be interested in the effect of returns and order book information of one asset on another asset. Hannan (1975) and Robinson (1975) are the earlier literature on using frequency domain to solve such problems. Mykland and Zhang (2006) discussed a general setup of the analysis of variance for continuous time regression.

3.5 Numerical Study: Comparison of estimators of co-volatility

Recall that our estmimator (at zero frequency $q = 0$) can be written as:

$$
\begin{aligned}
\sum_{\alpha} K_H(\lambda_\alpha) I(\lambda_\alpha) &:= \sum_{i,j=1}^{n} \Delta P_1(t_i)\Delta P_2(s_j) \sum_{\alpha} K_H(\lambda_\alpha) e^{-i(t_i - s_j)\alpha} \\
&=_{(i)} \sum_{i,j=1}^{n} \Delta P_1(t_i)\Delta P_2(s_j) k_H(t_i - s_j) \\
&=_{(ii)} \sum_{|h|<n} k\left(\frac{h}{H}\right)\gamma_h,
\end{aligned}
$$

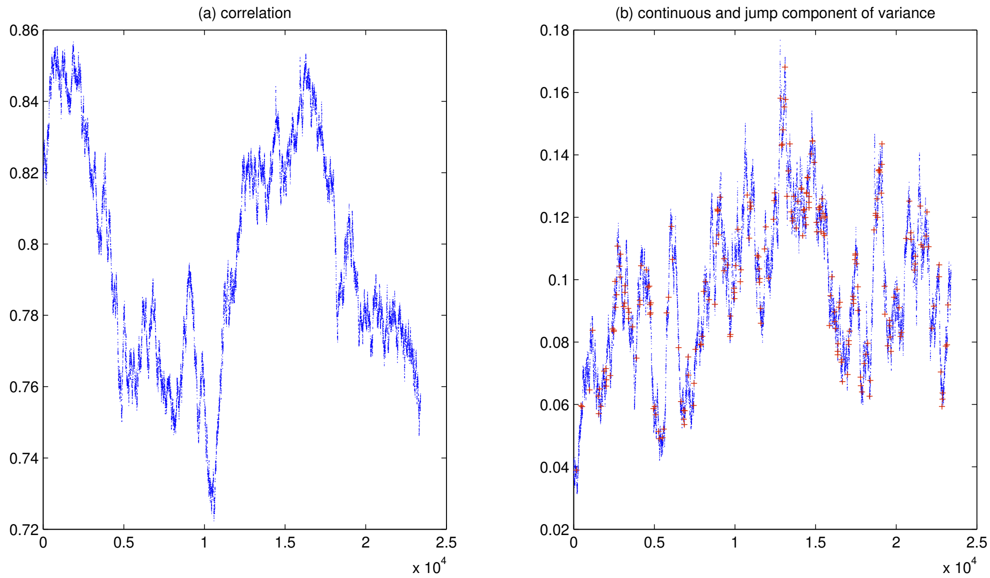
where (ii) holds only when the discretization points are synchronous and equally spaced. The form of the estimator we propose to implement is (i). In the theory sections above, we assumed no leverage effect; in the simulation studies however, we relax this assumption and see if our estimator is robust to the presence of the leverage. We consider two Data Generating Processes (DGP) for asset returns. For the first simulation, we consider the stochastic volatility model with a perfect leverage given in Barndorff-Nielsen et al. (2011). The volatility process is continuous whereas the instantaneous co-volatility is constant. For assets indexed by $j = 1, 2$,

$$
\begin{aligned}
dP_j(t) &= 0.03dt - 0.3\sigma_j(t)dB_j(t) + \sqrt{1-(0.3)^2}\sigma_j(t)dW(t) \qquad (3.24) \\
\sigma_j(t) &= \exp\{-5/16 + 1/8\varrho_j(t)\} \quad ; \quad d\varrho_j(t) = -1/40\varrho_j(t)dt + dB_j(t),
\end{aligned}
$$

and $\varrho_j(t)$ is initialized by $\varrho_j(0) \sim N(0, 20)$. The model implies that the covariance between the returns are $EdP_1(t)dP_2(t) = 0.91\sigma_1(t)\sigma_2(t)dt$. There is a perfect statistical leverage since a single Brownian motion $B_j(t)$ which is present in the return equation, drives the volatility process.

For the second simulation, the stochastic volatility is specified as a jump diffusion process and the instantaneous co-volatility coefficient follows the Cox-Ingersoll-Ross (CIR) process. This is a modification of the DGP considered in Aït-Sahalia et al. (2010) and

Figure 3.2: Simulated intraday instantaneous co-volatility and variance



Barndorff-Nielsen et al. (2004). For $j = 1, 2$, we have

$$\begin{aligned} dP_j(t) &= \sigma_j(t)dW_j(t) \\ d\sigma_j^2(t) &= \kappa_j\{\bar{\sigma}_j^2 - \sigma_j^2(t)\} + a_j\sigma_j(t)dB_j(t) + \sigma_j(t_-)J_j(t)dN_j(t), \end{aligned} \tag{3.25}$$

where the jump size follows $J_j(t) = \exp\{z_j(t)\}$ with $z_j(t) \sim N(\mu_j, s_j)$, and $N_j(t)$ is a poisson process with intensity λ_j. The leverage effect here is $EdW_j(t)dB_j(t) := \delta_j dt$, and the covariance between the Brownian motions that are present in the price equation is given by $EdB_1(t)dB_2(t) = \rho_t dt$. The parameter values taken from Aït-Sahalia et al. (2010). We let $\rho(t) = (e^{2x(t)} - 1)/(e^{2x(t)} + 1)$, where $x(t)$ follows the CIR process

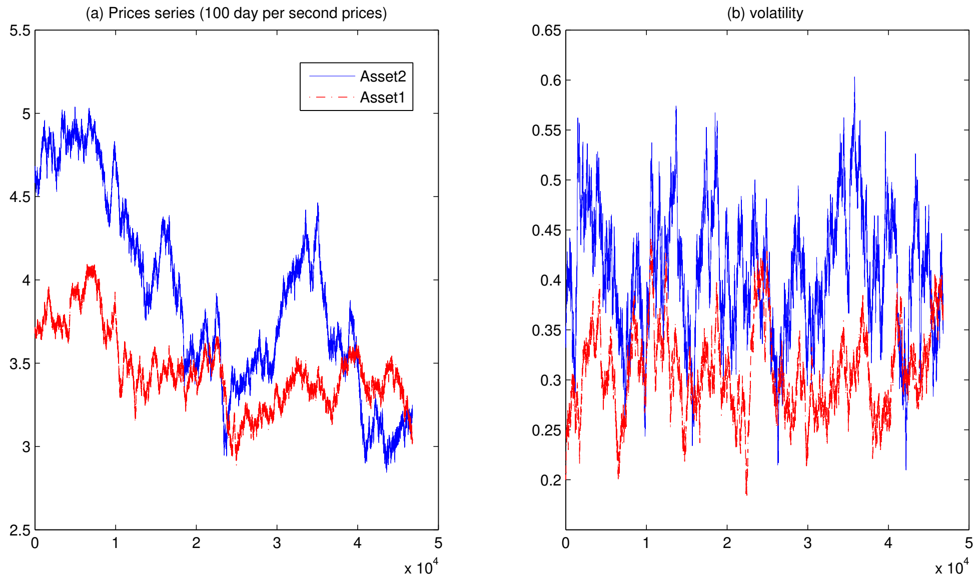
$$dx(t) = 0.03(0.64 - x(t))dt + 0.118x(t)dB(t).$$

Figure 3.2(a) shows the time series plot of ρ_t, and (b) shows $\sigma_1^2(t)$ decomposed into a continuous and a discontinuous components. Figure 3.3(a) shows the time series plot of $P_j(t), j = 1, 2$ and (b) shows $\sigma_j^2(t), j = 1, 2$. The DGP of the microstructure noise is formed with respect to the transaction time. We consider a correlated AR(1) noise processes with smoothly decaying cross autocovariances. This can be implemented by

$$U_j(t_{i,j}) = \bar{U}_j(t_{i,j}) + \varepsilon(t_{i,j}); \quad \bar{U}_j(t_{i,j}) = \alpha\bar{U}_j(t_{i-1,j}) + \epsilon_j(t_{i,j}), \tag{3.26}$$

where the idiosyncratic errors are independent Gaussian, i.e. $\epsilon_j(t_{i,j}) \sim NID(0, 1)$. The common disturbance that drives the correlation between two microstructure noise is sim-

Figure 3.3: Simulated price and variance - per second observation



ulated by

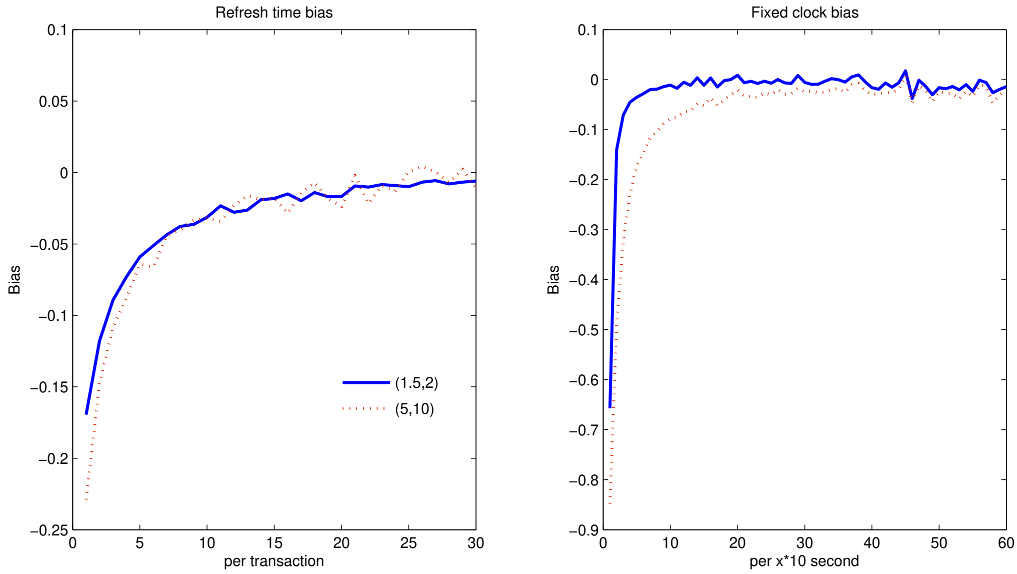
$$\varepsilon_l = 0.5\varepsilon_{l-1} + \xi_l, \text{ for } \{T_l\}_{1\leq l\leq N_T} = \{t_{i,1} \cup t_{j,2}, i = 1, \ldots, n_1, j = 1, \ldots, n_2\},$$

where $\xi_t \sim NID(0,1)$. Then we define $\{\varepsilon(t_{i,1})\}_{1\leq i\leq n_1}$ as $\{\varepsilon_l\}_{1\leq l\leq N_T}$ sampled at $\{T_l \cap t_{i,1}\}$ points. $\{\varepsilon(t_{j,2})\}_{1\leq j\leq n_2}$ is defined similarly. The variance of the noise is set to be proportionate to the sample integrated quarticity; $\zeta^2\sqrt{n_j{}^{-1}\sum_{i=1}^{n_j}\sigma_j^4(t_{i,j})}$, where $\zeta = \{0, 0.0.001, 0.01\}$ is the noise to signal ratio.

We simulate the one second data assuming 6.5 hours of daily trading time, yielding 23,400 daily data points over 100 Monte Carlo sample. The simulation is designed to assess the impact of the asynchronicity on the estimator: We Poisson sample the data at the rate of $\{(3/2, 30), (20, 30)\}$, where sampling at rate (a, b) means on average observation is made at every a and b seconds. To create a balanced sample for the rate $(3/2, 30)$, for the first asset, we sample on average at 1.5 second for the first half of the sample and at 30 second for the last half of the sample. For the second asset, we do this in reverse order. Then we have two assets that have the same number of transactions each day but traded very asynchronously.

Finally, we examine the properties of the estimators in higher dimensions. We consider a simple setting where the log prices are given by $P(t) = AB(t)$; $P(t)$ is the 10×1 vector of prices, $B(t)$ is the 3×1 independent Brownian motion and A is a factor loading matrix. This is poisson sampled at rate $\{2, 2, 4, 4, 8, 8, 10, 10, 30, 30\}$ and masked by i.i.d gaussian noise. Table 3.2 and 3.3 at the end of this chapter report the results for estimating the 2-dimensional covariation matrix, where the first asset is more often traded then the second asset. Table 3.4 reports the results for a higher dimension.

Figure 3.4: Covariation signature plot for simulated series - bias induced by data alignment for different sampling rates



(1) Realized Covariation: bias induced by data synchronization

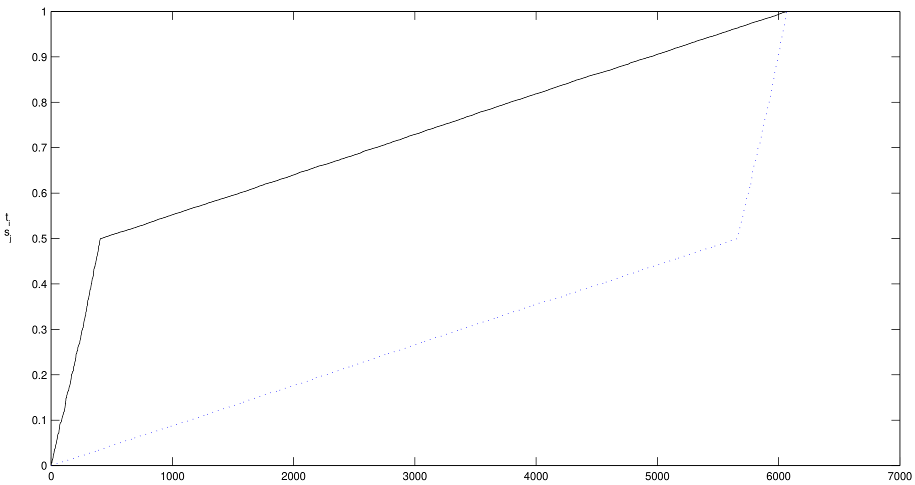
Table 3.1 reports the finite sample properties of the realized covariance (RC). The efficient price follows Brownian semimartingale, given in (3.24). The transaction time follows a homogenous poisson process, and the microstructure noise are correlated AR(1) processes given in (3.26). Asynchronous data is aligned using the 5 minute fixed clock time and the refresh time. The negative bias when there is no noise is consistent with the result of Proposition 3.2.

When microstructure noise is present, the variance estimate has a large positive bias. The sparse sampling (5 minute aligned data) is able to reduce such bias. However, the covariance estimate has a negative bias induced by the Epps effect which dominates the positive bias induced by the microstructure noise. The degree at which Epps effect dominates the noise effect depends on the degree of non-synchronicity. Figure 3.4 shows the covariation signature plot for the simulated series when the price is observed without the noise. It shows that given varying degrees of non-synchronicity (rate at which assets is traded), the higher frequency we align the data (moving leftwards in x axis) the more bias it induces in estimating the integrated covariance.

(2) Balanced Sample example

We elaborate further on the aforementioned sampling procedure: We first simulate the equally spaced data per one second for two assets. For the first asset, we sample on average at 1.5 second for the first half of the sample and at 30 second for the last half of the sample. For the second asset, this is done in reverse order - sample at 30 second for

Figure 3.5: Time stamp of two assets traded at opposite liquidity



the first half and 1.5 second for later part. See Figure 3.5. Then we have two assets that have the same number of transactions each day but traded very asynchronously. This is like a case where two assets have opposite liquidity profile over a day.

The sample size is $607,774$ over one hundred days and the refresh time aligned data reduces to a size around 750 per day. The large reduction in the sample size of the aligned data is due to severe non-synchronicity by simulation design. We compare the realized kernel and the proposed method over the range of bandwidths, $H = \{1, 5, 10, 20, 50, 100, \ldots, 750, 800\}$. Figure 3.6 shows that the proposed estimator is less sensitive to the choice of bandwidth - especially for large H. With large H, we can reduce the bias for the off-diagonal element more than we can do for the realized kernel. Our estimator is less sensitive to the choice of bandwidth for large values of H.

(3) Unbalanced Sample example

We carried out the same exercise as above but with the unbalanced sample sizes. We Poisson sample the data at rate $\{(3/2, 30), (3/2, 2), (20, 30)\}$. For example, sampling rate $(3/2, 2)$ means that we sample the first asset on average per 1.5 second and the second asset per 2 second. The first rate is to examine the effect of different liquidity and different sample sizes. The second and third rates are to examine the effect of sparse and intense sampling of asset prices of similar liquidity.

Figure 3.6 shows the results for sampling rate $(3/2, 30)$. The proposed estimator has a less bias and is less sensitive to the choice of the bandwidth for large values of H. When the noise is not present, the proposed method estimates the variance of more liquid asset more precisely. When the noise is present, the bandwidth should be large for the proposed estimator to perform better. The conclusion is similar for sampling rate $(20, 30)$ as shown in Figure 3.8. The improvement of using the Fourier domain estimator is most evident when estimating the variance of more liquid asset when two sample sizes are very different. The proposed covariation estimator performs better under large bandwidth. For sampling rate $(3/2, 2)$ in Figure 3.9, the difference of two estimator is less pronounced.

Each of these figures also show the accuracy of estimating the scalar function of the covariation matrix. We examined the maximum eigenvalues and the variance of portfolio with weight $[0.5, \sqrt{0.75}]$. Under the realistic noise to signal ratio and when two assets are of different liquidity, the proposed method delivers superior estimates. Regardless of sampling scheme, the proposed method does better in estimating these quantities when the effect of microstructure noise is not too dominant.

(4) Overall Comparison and Higher Dimension Case

Tables 3.2 and 3.3 show that the proposed estimator has the best bias profile. With carefully chosen bandwidth we can achieve the best root MSE under the presence of noise. When no noise is present, the Hayashi and Yoshida estimator performs well. The refresh time aligned method often performs better in estimating the integrated variance of the less traded asset; (2,2)th element. This is because it effectively aligns on the time stamp of less traded asset. As shown in the analysis of asymptotic bias, when there is no noise and the number of refresh time sample is smaller in size, then the realized kernel underperforms in terms of bias. The proposed estimator overall estimates the off-diagonal elements better. We observe also that the realized covariance estimator aligned on sparesly sampled data often performs well - this is because there are two opposing effects in terms of bias: negative bias from the Epps effect and positive bias from the microstructure noise. The advantage of our estimator is most clear when we estimate the covariance matrix of higher dimension, see Table 3.4. We estimate 10 dimensional integrated covariance matrix and compare the maximum of eigenvalues and variance of the equally weighted portfolio. Under no noise, the refresh time based method has a large bias. We calculate the optimal bandwidth as given in Theorem 3.1 for each element of covariance matrix and take their minimum, maximum and average them. Our estimator seems to yield a large variance, however it performs best under the optimal bandwidth.

(5) Concluding Remarks

In this numerical study section we have shown that the Fourier domain estimator performs well even under extreme asynchronicity and the presence of microstructure noise. It may be possible to make further improvements, given the literature on discrete time estimation of the spectral density. For example, Xiao and Linton (2002) and Hirukawa (2006) developed multiplicative bias reduction methods that can improve the performance under stronger conditions without sacrificing positive semi-definiteness.

3.6 Appendix: Proofs of the main results

We shall derive the results conditionally on the volatility matrix and the discretization time points, both of which are hence treated as deterministic throughout. In most cases, proofs are done under the most general framework; that is, when time stamps are asynchronous and sample sizes are unbalanced. The transaction time of the first and second assets is denoted either by $\{t_{i,1}, t_{j,2}\}$ or $\{t_i, s_j\}$.

Lemma 3.1. *Suppose $P(t)$ defined on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ satisfies Assumption 3.1 with $\mu(t) = 0$. Let $f(t, s; q)$ be a bounded, measurable and square-integrable function. Then, it follows that for any $a, b, c, d = 1, \ldots, p$,*

$$E\left[\int_0^{2\pi}\int_0^{2\pi} f(t,s;q)dP_a(s)dP_b(t), \int_0^{2\pi}\int_0^{2\pi} f(t,s;q')dP_c(s)dP_d(t)\right] \tag{3.27}$$
$$= \int_0^{2\pi}\int_0^{2\pi} f(t,s;q)f(t,s;q')\ d[P_a,P_c](s)d[P_b,P_d](t)$$
$$+ \int_0^{2\pi}\int_0^{2\pi} f(t,s;q)f(s,t;q')\ d[P_a,P_d](s)d[P_b,P_c](t),$$

where the double stochastic integral is in Wiener-Itô sense.

Proof of Lemma 3.1. From the definition of the double Wiener-Itô integral and standard properties of symmetrization operation, it is straightforward to see that

$$\int_0^{2\pi}\int_0^{2\pi} f(t,s;q)dP_a(s)dP_b(t)$$
$$= \int_0^{2\pi}\int_0^{t} f(t,s;q)dP_a(s)dP_b(t) + \int_0^{2\pi}\int_0^{t} f(s,t;q)dP_b(s)dP_a(t).$$

The integrand here is measurable with respect to $\mathcal{F}_t$, and two terms above are martingale. Consequently, (3.27) can be expressed as

$$\left[\begin{array}{c} \int_0^{2\pi}\int_0^t f(t,s;q)dP_a(s)dP_b(t) + \int_0^{2\pi}\int_0^t f(s,t;q)dP_b(s)dP_a(t), \\ \int_0^{2\pi}\int_0^t f(t,s;q')dP_c(s)dP_d(t) + \int_0^{2\pi}\int_0^t f(s,t;q')dP_d(s)dP_c(t) \end{array}\right].$$

Then one of the cross terms is given by

$$
\begin{aligned}
& E\Big[\int_0^{2\pi}\int_0^t f(t,s;q)dP_a(s)dP_b(t), \int_0^{2\pi}\int_0^t f(s,t;q')dP_d(s)dP_c(t)\Big] \qquad (3.28)\\
&= E\int_0^{2\pi}\Big(\int_0^t f(t,s;q)dP_a(s)\Big)\Big(\int_0^t f(s,t;q')dP_d(s)\Big)\, d[P_b,P_c](t).\\
&= \int_0^{2\pi} E\Big(\int_0^t f(t,s;q)dP_a(s)\Big)\Big(\int_0^t f(s,t;q')dP_d(s)\Big)\, d[P_b,P_c](t).\\
&= \int_0^{2\pi}\int_0^t f(t,s;q)f(s,t;q')\; d[P_a,P_d](s)d[P_b,P_c](t). \qquad (3.29)
\end{aligned}
$$

Interchange of the expectation operator and the integral is justified by Fubini's Theorem. Note that $d[P_b,P_c](t) = [P_b,P_c]'(t)dt$, where the prime denotes the time derivative. The same arguments apply to the remaining terms, completing the proof. ■

Lemma 3.2. *Consider the following kernel-weighted off-diagonal step functions:*

$$
\begin{aligned}
f_n(t,s;q) &= \sum_{i\neq j} e^{-\mathrm{i}qt_j} k_H(t_i-t_j) 1_{[t_{i-1},t_i[}(t) 1_{[t_{j-1},t_j[}(s) \qquad (3.30)\\
g_n(t,s;q) &= \sum_{i=1}^n\sum_{j=1}^n e^{-\mathrm{i}qt_j} k_H(t_i-t_j) 1_{[t_{i-1},t_i[}(t) 1_{[t_{j-1},t_j[}(s)\ ,
\end{aligned}
$$

where $\{t_i; i=1,\dots,n\}$ are the discretization points that satisfy Assumption 3.2, and the kernel $k_H(t_i-t_j) =: k\left((t_i-t_j)/(H\overline{\Delta} t)\right)$ where $k(\cdot)$ is a lag window satisfying 3.3. Then

$$
\begin{aligned}
& \frac{n}{H}\int_0^{2\pi}\int_0^{2\pi}\Big\{f_n^2(t,s;q) + f_n(t,s;q)f_n(s,t;q)\Big\}\; d[P,P](t)d[P,P](s) \qquad (3.31)\\
& \quad \longrightarrow 2\|k\|^2\int_0^{2\pi} e^{-\mathrm{i}2tq}([P,P]'(t))^2 d\mathcal{Q}(t),
\end{aligned}
$$

where $\mathcal{Q}(t)$ is the limit of the empirical quadratic covariation $\mathcal{Q}^{(n)}(t) = n\sum_{i:t_i<t}(\Delta t_i)^2$.

Proof of Lemma 3.2. We first show f_n can be replaced by g_n in the integrand with an error of order $o(1)$. Given a single asset observed at discretizing points $\{t_i\}$ we have

$$
\begin{aligned}
& \int_0^{2\pi}\int_0^{2\pi}\big[g_n(t,s;q) - f_n(t,s;q)\big]\; dsdt\\
&= \int_0^{2\pi}\int_0^{2\pi}\Big\{\sum_{i=1}^n\sum_{j=1}^n - \sum_{i\neq j}\Big\}\cdot e^{-\mathrm{i}t_jq}k_H(t_i-t_j)\cdot 1_{[t_{i-1},t_i[}(t)1_{[t_{j-1},t_j[}(s)dsdt\\
&= \int_0^{2\pi}\int_0^{2\pi}\Big\{\sum_{j=1}^n e^{-\mathrm{i}t_jq}1_{[t_{j-1},t_j[}(t)1_{[t_{j-1},t_j[}(s)\Big\}dsdt\ \le\ C\Big(\sup_{1\le j\le n}\Delta t_j\Big) = O\Big(\frac{1}{n}\Big), \qquad (3.32)
\end{aligned}
$$

by Assumption 3.2. Generalising this idea we can rewrite (3.31) as

$$\begin{aligned}
&\frac{n}{H}\int_0^{2\pi}\int_0^{2\pi}\left\{g_n^2(t,s;q)+g_n(t,s;q)g_n(s,t;q)\right\}\,d[P,P](t)d[P,P](s)+o(1)\\
=&\frac{n}{H}\sum_{h=0}^{n-1}\sum_{j=1}^{n-h}\Big[(e^{-\mathrm{i}t_j2q}+e^{-\mathrm{i}t_jq}e^{-\mathrm{i}t_{j+h}q})k_H^2(t_{j+h}-t_j)[P,P]'(t_{j+h})[P,P]'(t_j)\Delta t_{j+h}\Delta t_j\\
&+\sum_{h=1}^{n-1}\sum_{j=1+h}^{n}(e^{-\mathrm{i}t_j2q}+e^{-\mathrm{i}t_jq}e^{-\mathrm{i}t_{j-h}q})k_H^2(t_{j-h}-t_j)[P,P]'(t_{j-h})[P,P]'(t_j)\Delta t_{j-h}\Delta t_j\Big]+o(1)\\
=&\frac{n}{H}\Bigg[\sum_{h=0}^{n-1}k^2\left(\frac{t_h/\overline{\Delta t}}{H}\right)\sum_{j=1}^{n-h}(e^{-\mathrm{i}t_j2q}+e^{-\mathrm{i}t_jq}e^{-\mathrm{i}t_{j+h}q})[P,P]'(t_{j+h})[P,P]'(t_j)\Delta t_{j+h}\Delta t_j\\
&+\sum_{h=1}^{n-1}k^2\left(\frac{-t_h/\overline{\Delta t}}{H}\right)\sum_{j=1+h}^{n}(e^{-\mathrm{i}t_j2q}+e^{-\mathrm{i}t_jq}e^{-\mathrm{i}t_{j-h}q})[P,P]'(t_{j-h})[P,P]'(t_j)\Delta t_{j-h}\Delta t_j\Bigg]+o(1)
\end{aligned}\tag{3.33}$$

Further,

$$\begin{aligned}
(3.33) &= \frac{1}{H}\left[\sum_{h=0}^{n-1}k^2\left(\frac{t_h/\overline{\Delta t}}{H}\right)+\sum_{h=1}^{n-1}k^2\left(\frac{-t_h/\overline{\Delta t}}{H}\right)+1-1\right]\\
&\times n\sum_{j=1}^{n}(e^{-\mathrm{i}t_j2q}+e^{-\mathrm{i}t_jq}e^{-\mathrm{i}t_{j+h}q})[P,P]'(t_{j+h})[P,P]'(t_j)\Delta t_{j+h}\Delta t_j+o(1)+O\left(\frac{1}{n}\right) &(3.34)\\
&\longrightarrow 2\|k\|^2\int_0^{2\pi}e^{-\mathrm{i}t2q}([P,P]'(t))^2d\mathcal{Q}(t). &(3.35)
\end{aligned}$$

It is straightforward to see that the error in approximating the kernel $k_H(t_i-t_{i-h})$ by $k_H(t_h)$ in (3.33) is negligible. Taylor expanding the lag window we have

$$k_H(t_i-t_{i-h})-k_H(t_h)\simeq\frac{t_i-t_{i-h}-t_h}{\overline{\Delta t}H}k'\left(\frac{t_h}{\overline{\Delta t}H}\right)+\frac{1}{2}\left(\frac{t_i-t_{i-h}-t_h}{\overline{\Delta t}H}\right)^2k''\left(\frac{t_h}{\overline{\Delta t}H}\right),$$

where

$$\frac{t_i-t_{i-h}-t_h}{\overline{\Delta t}H}=\sum_{l=1}^{h}(\Delta t_{l+i-h}-\Delta t_l)\frac{1}{\overline{\Delta t}H}\le 2h\left(\frac{\sup_l\Delta t_l}{\overline{\Delta t}H}\right)=Ch/H.$$

Then the approximating error in (3.33) is bounded by

$$C\sup_{j,h}(\Delta t_{j-h}\Delta t_j)\frac{n}{H}\sum_{|h|<n}\sum_{j=1+h}^{n}\left\{\frac{h}{H}k'\left(\frac{h}{H}\right)+\left(\frac{h}{H}\right)^2k''\left(\frac{h}{H}\right)\right\}\simeq C\frac{1}{H}\int_{-\infty}^{\infty}x^2k''(x)dx,$$

by Assumptions 3.2 and 3.3.

Further, the approximation error from (3.33) to (3.34) is given by

$$
\begin{aligned}
&\frac{n}{H}\left[\sum_{h=0}^{n-1} k^2\left(\frac{t_h/\overline{\Delta t}}{H}\right)+\sum_{h=1}^{n-1} k^2\left(\frac{-t_h/\overline{\Delta t}}{H}\right)\right] \\
&\qquad\qquad \times 2\left(\sum_{j=1}^{h}(e^{-\mathrm{i}t_j 2q}+e^{-\mathrm{i}t_j q}e^{-\mathrm{i}t_{j+h}q})[P,P]'(t_{j+h})[P,P]'(t_j)\Delta t_{j+h}\Delta t_j\right) \\
&\qquad \leq C(\sup \Delta t_j)^2 n \sum_{|h|<n}^{n-1} \frac{h}{H}k^2\left(\frac{h}{H}\right) \simeq \frac{1}{n}C\int_{-\infty}^{\infty} xk^2(x)dx.
\end{aligned}
$$

Lastly, (3.35) is justified by the fact that both $\frac{n}{H}\sum_{|h|<n} k^2\left(\frac{h}{H}\right)\sum_{j=1}^{n}\Delta t_{j+h}\Delta t_j$ and $\frac{n}{H}\sum_{|h|<n} k^2\left(\frac{h}{H}\right)\sum_{j=1}^{n}\Delta t_j\Delta t_j$ approach to the same limit due to the presence of the kernel weights, and by Riemann approximation of the integral $\int_{-\infty}^{\infty} k^2(x)dx$. ■

Lemma 3.3. *Consider the following off-diagonal step functions*

$$
\begin{aligned}
f_n(t,s;q) &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} e^{-\mathrm{i}s_j q}e^{-\mathrm{i}(t_i-s_j)\alpha}1_{[t_{i-1},t_i[}(t)1_{[s_{j-1},s_j[}(s)1_{\{I_{i,1}\cap I_{j,2}=\emptyset\}}(t,s) \\
g_n(t,s;q) &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} e^{-\mathrm{i}s_j q}e^{-\mathrm{i}(t_i-s_j)\alpha}1_{[t_{i-1},t_i[}(t)1_{[s_{j-1},s_j[}(s),
\end{aligned}
$$

where the discretization points $\{t_i, s_j\}$ satisfy Assumption 3.2. Then we have

$$
\int_0^{2\pi}\int_0^{2\pi} f_n(t,s;q)dsdt = \int_0^{2\pi}\int_0^{2\pi} g_n(t,s;q)dsdt + O\left(\frac{1}{n_1\wedge n_2}\right). \tag{3.36}
$$

Proof of Lemma 3.3. The arguments closely follow those in (3.32). The only difference is that here we consider two different assets with discretizing points $\{t_i, s_j\}$.

The difference between the integrals in (3.36) is given by

$$
\begin{aligned}
&\int_0^{2\pi}\int_0^{2\pi}\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left[1-1_{\{I_{i,1}\cap I_{j,2}=\emptyset\}}(t,s)\right]e^{-\mathrm{i}qs_j}e^{-\mathrm{i}\alpha(t_i-s_j)}1_{[t_{i-1},t_i[}(t)1_{[s_{j-1},s_j[}(s)\right)dsdt \\
&= \int_0^{2\pi}\int_0^{2\pi}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} e^{-\mathrm{i}qs_j}e^{-\mathrm{i}\alpha(t_i-s_j)}1_{[t_{i-1},t_i[}(t)1_{[s_{j-1},s_j[}(s)1_{\{I_{i,1}\cap I_{j,2}\neq\emptyset\}}(t,s)dsdt \\
&\leq \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\Delta t_i\Delta s_j 1_{\{I_{i,1}\cap I_{j,2}\neq\emptyset\}} \leq \sup_{1\leq i\leq n_1}(\Delta t_i)\sup_{1\leq j\leq n_2}(\Delta s_j)\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}1_{\{I_{i,1}\cap I_{j,2}\neq\emptyset\}} \\
&= \frac{(s^*)^2}{n_1 n_2}\#\{t_i\cup s_j, 0\leq t_i, s_j\leq 2\pi\} = O\left(\frac{1}{n_1\wedge n_2}\right),
\end{aligned}
$$

because the total number of union of time two stamps is of order $O(n_1\vee n_2)$. ■

Lemma 3.4. *Suppose that $P(t)$ is defined on some probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ and satisfies Assumption 3.1, and denote by $\mathcal{G} = \sigma(P)$ a sub-σ-field of $\mathcal{F}$. Let Z be a standard Gaussian variable on a suitable extension of the probability space, and $\mathcal{V}$ be the $\mathcal{G}$-measurable stochastic variance. Then for $f_n(\cdot)$ given in Lemma 3.2,*

$$\sqrt{\frac{n}{H}} \int_0^{2\pi} \int_0^{2\pi} f_n(t,s;q) dP_1(s) dP_2(t) \Longrightarrow^{stably} \sqrt{\mathcal{V}} Z,$$

provided that the conditions of Theorem 3.2 of Jacod (1997) are met.

Proof of Lemma 3.4. Given the discretized filtration $\mathcal{F}_i$, $i = \max_j\{t_j \leq t\}$ and the discretized sequence adapted to $\mathcal{F}_i$:

$$\chi_i^n = \sqrt{n/H} \Delta P_1(t_i) \sum_{j:s_j<t_i} \Delta P_2(s_j) k_H(t_i - s_j) e^{-\mathrm{i} s_j q},$$

we show stable convergence of $Z_t^n := \sum_{\max_i\{t_i\leq t\}} \chi_i^n$ to $Z_t = \int_0^t v_s dW_s$, a $\mathcal{F}_t$-conditional Gaussian martingale. Then by Theorem 3.2 of Jacod (1997) we have $Z^n \Longrightarrow Z$ stably. For future reference we note that a sufficient condition for the conditional Lindeberg condition in (4) is the Lyapunov condition $\sum_i E(\{\chi_i^n\}^{2+\varepsilon} | \mathcal{F}_{i-1}) \to_p 0$, for some $\varepsilon > 0$. We will show this for $\varepsilon = 2$ in the proof of Theorem 3.1 later. ■

3.6.1 Proof of Proposition 3.1

Proof. For the sake of simplicity of notation take $a = 1$ and $b = 2$, w.l.o.g. Similar to (3.15),

$$\begin{aligned}
&\sum_{|\alpha|\leq m/2} K_H(\lambda_\alpha) \big[\mathcal{F}_n(dU_1)(\alpha)\big] \big[\mathcal{F}_n(dU_2)(q-\alpha)\big] \\
&\qquad = \sum_{|\alpha|\leq m/2} K_H(\lambda_\alpha) \left[\sum_{i=1}^{n_1} e^{-\mathrm{i}\alpha t_i} \Delta U_1(t_i)\right] \left[\sum_{j=1}^{n_2} e^{-\mathrm{i}(q-\alpha)s_j} \Delta U_2(s_j)\right] \\
&\qquad = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Delta U_1(t_i) \Delta U_2(s_j) e^{-\mathrm{i}q s_j} k_H(t_i - s_j).
\end{aligned} \tag{3.37}$$

We first consider the terms that are not affected by the end points (i.e. t_{n_1}, t_0, s_{n_2} and s_0). Expanding the difference operators we see that those terms are given by

$$\begin{aligned}
\sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} U_1(t_i) U_2(s_j) \Big\{ & e^{-\mathrm{i}q s_j} \Big[-k_H\big(t_{i+1} - s_j\big) + k_H\left(t_i - s_j\right)\Big] \\
& + e^{-\mathrm{i}q s_{j+1}} \Big[k_H(t_{i+1} - s_{j+1}) - k_H(t_i - s_{j+1})\Big]\Big\}.
\end{aligned} \tag{3.38}$$

By repeated use of Taylor expansion the terms in the curly bracket of (3.38) become

$$\Bigg\{e^{-\mathrm{i}s_jq}\Bigg[-k\left(\frac{t_i-s_j+\Delta t_{i+1}}{\bar{H}}\right)+k\left(\frac{t_i-s_j}{\bar{H}}\right)\Bigg]$$

$$+e^{-\mathrm{i}s_{j+1}q}\Bigg[k\left(\frac{t_i-s_j+\Delta t_{i+1}-\Delta s_{j+1}}{\bar{H}}\right)-k\left(\frac{t_i-s_j-\Delta s_{j+1}}{\bar{H}}\right)\Bigg]\Bigg\}$$

$$=\Bigg\{-e^{-\mathrm{i}s_jq}\left[\frac{\Delta t_{i+1}}{\bar{H}}k'\left(\frac{t_i-s_j}{\bar{H}}\right)\right]+e^{-\mathrm{i}s_{j+1}q}\left[\frac{\Delta t_{i+1}}{\bar{H}}k'\left(\frac{t_i-s_j-\Delta s_{j+1}}{\bar{H}}\right)\right]+o(1)\Bigg\}$$

$$=\Bigg\{-e^{-\mathrm{i}s_jq}\left[\frac{\Delta t_{i+1}}{\bar{H}}k'\left(\frac{t_i-s_j}{\bar{H}}\right)-\frac{\Delta t_{i+1}}{\bar{H}}k'\left(\frac{t_i-s_j-\Delta s_{j+1}}{\bar{H}}\right)\right]$$

$$+e^{-\mathrm{i}s_{j+1}q}\left[\frac{\Delta t_{i+1}}{\bar{H}}k'\left(\frac{t_i-s_j-\Delta s_{j+1}}{\bar{H}}\right)\right]-e^{-\mathrm{i}s_jq}\frac{\Delta t_{i+1}}{\bar{H}}k'\left(\frac{t_i-s_j-\Delta s_{j+1}}{\bar{H}}\right)+o(1)\Bigg\}$$

$$=\Bigg\{e^{-\mathrm{i}s_jq}\left[\frac{\Delta t_{i+1}\Delta s_{j+1}}{\bar{H}^2}k''\left(\frac{t_i-s_j}{\bar{H}}\right)\right]+e^{-\mathrm{i}s_jq}\left[(e^{-\mathrm{i}\Delta s_{j+1}q}-1)\frac{\Delta t_{i+1}}{\bar{H}}k'\left(\frac{t_i-s_j}{\bar{H}}\right)\right]+o(1)\Bigg\}$$

where $\bar{H}=H\widetilde{\Delta t}_{12}$.

In the sequel $\widetilde{\Delta t}$ is taken to mean $\widetilde{\Delta t}_{12}$ for the sake of simplicity. Note that under Assumptions 3.2 and 3.4 $\Delta t_{i+1}\Delta s_{j+1}\frac{1}{\bar{H}^2}=o(1)$ and $\Delta t_{i+1}\frac{1}{\bar{H}}=O(H^{-1})=o(1)$.

When samples are equally spaced and balanced, the curely bracket terms can be simplified further as follows:

$$\Bigg\{\cdot\Bigg\}=-e^{-\mathrm{i}s_jq}\frac{1}{H^2}k''\left(\frac{i-j}{H}\right)+e^{-\mathrm{i}s_jq}\big(e^{-\mathrm{i}\Delta s_{j+1}q}-1\big)\frac{1}{H}k'_H\left(\frac{i-j-1}{H}\right).$$

It now follows that (3.38) can be written as a sum of the following two terms (up to additional terms of negligible order $o(1)$):

$$A_1+A_2=\frac{1}{\bar{H}^2}\sum_{i=1}^{n_1-1}\sum_{j=1}^{n_2-1}e^{-\mathrm{i}s_jq}\left[U_1(t_i)U_2(s_j)k''\left(\frac{t_i-s_j}{\bar{H}}\right)\right]\Delta t_{i+1}\Delta s_{j+1} \tag{3.39}$$

$$+\ \frac{1}{\bar{H}}\sum_{i=1}^{n_1-1}\sum_{j=1}^{n_2-1}e^{-\mathrm{i}s_jq}\cdot\big(e^{-\mathrm{i}\Delta s_{j+1}q}-1\big)\left[U_1(t_i)U_2(s_j)k'\left(\frac{t_i-s_j}{\bar{H}}\right)\right]\Delta t_{i+1}. \tag{3.40}$$

We show that A_2 is of negligible order of magnitude, and that A_1 is the leading term that dominates in large sample. The expectation of A_2 is bounded above by:

$$
\begin{aligned}
E(A_2) &\leq \frac{1}{\widetilde{\Delta t} H} \sup_j \left|1 - e^{-\mathrm{i}q\Delta s_{(j+1)}}\right| \sup_i \left(\Delta t_{i+1}\right) \left| \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} E\big[U_1(t_i)U_2(s_j)\big] \cdot k'\left(\frac{t_i - s_j}{\bar{H}}\right) \right| \\
&= C \frac{n_1 \vee n_2}{H} \frac{1}{n_1 n_2} \left| \left[\sum_{|t_i - s_j|/\widetilde{\Delta t} \leq \sqrt{H}} + \sum_{|t_i - s_j|/\widetilde{\Delta t} > \sqrt{H}} \right] E\big[U_1(t_i)U_2(s_j)\big] \cdot k'\left(\frac{t_i - s_j}{\bar{H}}\right) \right| \\
&\leq \frac{C}{H(n_1 \wedge n_2)} \Bigg\{ \sup_{|t_i - s_j|/\widetilde{\Delta t} \leq \sqrt{H}} \left| k'\left(\frac{t_i - s_j}{\bar{H}}\right) \right| \cdot \left| \sum_{|t_i - s_j|/\widetilde{\Delta t} \leq \sqrt{H}} \gamma\left(\frac{t_i - s_j}{\widetilde{\Delta t}}\right) \right| \\
&\qquad + \sup_{|t_i - s_j|/\widetilde{\Delta t} > \sqrt{H}} \left| \gamma\left(\frac{t_i - s_j}{\widetilde{\Delta t}}\right) \right| \cdot \left| \sum_{|t_i - s_j|/\widetilde{\Delta t} > \sqrt{H}} k'\left(\frac{t_i - s_j}{\bar{H}}\right) \right| \Bigg\} \\
&= \frac{C}{H(n_1 \wedge n_2)} \Big\{ O(1) O(n_1 \wedge n_2) + o(1) \Big\} \\
&= O\left(\frac{1}{H}\right) = o(1),
\end{aligned}
$$

where in the second and the last equality we used Taylor approximation and the fact that $k'(0) = 1$ along with Assumption 3.6, respectively.

Now consider a set $\mathcal{S} := \{i, j, r, l; (t_i - t_r)/\overline{\Delta} t < H^\mu, (s_j - s_l)/\overline{\Delta} s < H^\mu\}$ where μ is some positive constant between $(0, 1)$. Note that $\max_{i,r}(t_i - t_r)/\overline{\Delta} t = n$, and recall $H = n^\kappa$, $\kappa \in (0, 1)$.

We then see that $E(A_2^2)$ is bounded above by

$$
\begin{aligned}
&\frac{CH^{-2}}{(n_1 \wedge n_2)^2} \left\{ \sum_{i,j,r,l} E\Big[U_1(t_i)U_2(s_j)U_1(t_r)U_2(s_l)\Big] k'\left(\frac{t_i - s_j}{\bar{H}}\right) k'\left(\frac{t_r - s_l}{\bar{H}}\right) \right\} \qquad (3.41) \\
&= \frac{CH^{-2}}{(n_1 \wedge n_2)^2} \left\{ \sum_{i,j,r,l \in \mathcal{S}} + \sum_{i,j,r,l \in \mathcal{S}^c} \right\} E\Big[U_1(t_i)U_2(s_j)U_1(t_r)U_2(s_l)\Big] k'\left(\frac{t_i - s_j}{\bar{H}}\right) k'\left(\frac{t_r - s_l}{\bar{H}}\right) \\
&\leq \frac{CH^{-2}}{(n_1 \wedge n_2)^2} \sup_{i,j,r,l \in \mathcal{S}} \left| EU_1(t_i)U_2(s_j)U_1(t_r)U_2(s_l) \right| \\
&\qquad \times \left| \sum_{i,j} \sum_{|h|,|v| < H^\mu} k'\left(\frac{t_i - s_j}{\bar{H}}\right) k'\left(\frac{t_{i-h} - s_{j-v}}{\bar{H}}\right) \right| \\
&\quad + \frac{CH^{-2}}{(n_1 \wedge n_2)^2} n_1^2 n_2^2 \sup_{i,j,r,l \in \mathcal{S}^c} \left| EU_1(t_i)U_2(s_{i-h})U_1(t_r)U_2(s_{r-v}) \right| = B_1 + B_2.
\end{aligned}
$$

When the data is balanced and equally spaced, B_1 simplifies to

$$\frac{C}{(n_1\wedge n_2)^2H^2}\Bigg[\sup_{|i-r|<H^\mu,|h-v|<H^\mu}\Big|EU_1(t_i)U_2(s_{i-h})U_1(t_r)U_2(s_{r-v})\Big| \\ \times\Bigg|\sum_{|i-r|<H^\mu,|h-v|<H^\mu}k'\left(\frac{t_i-s_{i-h}}{\bar{H}}\right)k'\left(\frac{t_r-s_{r-v}}{\bar{H}}\right)\Bigg|\Bigg]. \tag{3.42}$$

In particular, when the sample size of two assets is equal, it holds that $(t_i-s_{i-h})/\bar{H}=h/H+o(1)$ under Assumption 3.2. Hence the squared bracket term in (3.42) is given by

$$\begin{aligned}\sum_{|i-r|<H^\mu}\sum_{|h-v|<H^\mu}k'\left(\frac{h}{H}\right)k'\left(\frac{v}{H}\right) &= 2H^\mu n\sum_{|h-v|<H^\mu}k'\left(\frac{h}{H}\right)k'\left(\frac{v}{H}\right)\\ &= 2H^\mu n\left\{\sum_{0\le l<H^\mu}\sum_{h=1+l}^{n}k'\left(\frac{h}{H}\right)k'\left(\frac{h-l}{H}\right)+\sum_{0<l<H^\mu}\sum_{h=1}^{n-l}k'\left(\frac{h}{H}\right)k'\left(\frac{h+l}{H}\right)\right\}\\ &\le 4H^{2\mu}n\sum_{h=1}^{n}\left\{k'\left(\frac{h}{H}\right)\right\}^2.\end{aligned}$$

In the unbalanced case, the Riemann approxmination gives

$$\sum_{i=1}^{n_1-1}\sum_{j=1}^{n_2-1}\left\{k'\left(\frac{t_i-s_j}{\bar{H}}\right)\right\}^2=\big(n_1\wedge n_2\big)H\int_{-\infty}^{\infty}\{k'(x)\}^2dx+o(1),$$

and the order of $\#\{0\le i,r\le n_1;\frac{t_i-t_r}{\widetilde{\Delta t}}<H^\mu\}$ is the same as when the data is equally spaced under Assumption 3.2; therefore, it follows that

$$B_1=\frac{CH^{-2}}{(n_1\wedge n_2)^2}\times\left[4\rho(0)(n_1\wedge n_2)H^{2\mu+1}\int_{-\infty}^{\infty}\Big(k'(x)\Big)^2dx\right]=O\left(\frac{H^{2\mu-1}}{n_1\wedge n_2}\right)=o(1).$$

Also, we have $B_2=CH^{-2}(n_1\wedge n_2)^{-2}n_1^2n_2^2\cdot\sup_{|\tau|>H^\mu}\rho(\tau)$ which vanishes in large sample as a consequence of (17) in Assumption 3.6.

We next move on to A_1. Its expectation is given by

$$\bar{H}^{-2}\sum_{|t_i-s_j|/\widetilde{\Delta t}\le\sqrt{H}}E\Big[U_1(t_i)U_2(s_j)\Big]e^{-\mathrm{i}s_jq}k''\left(\frac{t_i-s_j}{\bar{H}}\right)\Delta t_{i+1}\Delta s_{j+1} \\ +\bar{H}^{-2}\sum_{|t_i-s_j|/\widetilde{\Delta t}>\sqrt{H}}E\Big[U_1(t_i)U_2(s_j)\Big]e^{-\mathrm{i}s_jq}k''\left(\frac{t_i-s_j}{\bar{H}}\right)\Delta t_{i+1}\Delta s_{j+1}\;=\;D_1+D_2,$$

where D_2 is bounded above by

$$\begin{aligned}
&\bar{H}^{-2}\sup_i(\Delta t_{i+1})\sup_j(\Delta s_{j+1})\sup_{|t_i-s_j|/\widetilde{\Delta t}>\sqrt{H}}\left|EU_1(t_i)U_2(s_j)\right|\sum_{|t_i-s_j|/\widetilde{\Delta t}>\sqrt{H}}\left|k''\left(\frac{t_i-s_j}{\bar{H}}\right)\right| \\
\le\;& C\left(\frac{n_1\vee n_2}{H^2}\right)\sup_{|t_i-s_j|/\widetilde{\Delta t}>\sqrt{H}}\left|\gamma\left(\frac{|t_i-s_j|}{\widetilde{\Delta t}}\right)\right|\cdot H\int_{-\infty}^{\infty}|k''(x)|\,dx=o(1),
\end{aligned}$$

and moreover

$$\begin{aligned}
D_1 &\le \frac{(n_1\vee n_2)^2}{(2\pi)^2H^2}\frac{s^*}{n_1}\frac{s^*}{n_2}\sum_{|t_i-s_j|/\widetilde{\Delta t}\le\sqrt{H}}E\Big[U_1(t_i)U_2(s_j)\Big]k''\left(\frac{t_i-s_j}{\bar{H}}\right)e^{-\mathrm{i}s_jq} \\
&=\left(\frac{s^*}{2\pi}\right)^2\frac{n_1\vee n_2}{H^2(n_1\wedge n_2)}|k''(0)|\sum_{|t_i-s_j|/\widetilde{\Delta t}\le\sqrt{H}}E\Big[U_1(t_i)U_2(s_j)\Big]e^{-\mathrm{i}s_jq}+o(1) \\
&\le\left(\frac{s^*}{2\pi}\right)^2\frac{n_1\vee n_2}{H^2}|k''(0)|\,\Gamma_{12}+o(1)=O\left(\frac{n_1\vee n_2}{H^2}\right)
\end{aligned}$$

by Assumption 3.6. The expectation of the square of (3.39) is bounded above by

$$\begin{aligned}
&C\left(\frac{n_1\vee n_2}{H^2(n_1\wedge n_2)}\right)^2E\left\{\sum_{i=1}^{n_1-1}\sum_{j=1}^{n_2-1}\Big[U_1(t_i)U_2(s_j)\Big]k''\left(\frac{t_i-s_j}{\bar{H}}\right)\right\}^2 \\
&=C\left(\frac{n_1\vee n_2}{H^2(n_1\wedge n_2)}\right)^2\rho(0)\cdot 4(n_1\wedge n_2)H^{2\mu+1}\int_{-\infty}^{\infty}\big(k''(x)\big)^2dx+o(1) \\
&=O\left(\frac{(n_1\vee n_2)^2}{n_1\wedge n_2}H^{2\mu-3}\right). \qquad (3.43)
\end{aligned}$$

All other terms involving the end points are of smaller order by similar arguments given in Barndorff-Nielsen et al. (2011) along with Assumption 3.7. The proof is complete. ■

3.6.2 Proof of Proposition 3.2 and Theorem 3.1

PROOF. The proof consists of three parts; we show stable convergence of the diagonal and off-diagonal terms of the centered estimator in the first and second parts, respectively, followed by a brief justification for joint convergence via the Cramér-Wold device.

Without loss of generality consider the first element of the centred estimator (without noise); that is, the term involving the price process of the first asset:

$$\begin{aligned}
\mathcal{E}_1 &= \Big[\widehat{\mathcal{F}}(\Sigma^P)(q)-\mathcal{F}(\Sigma^P)(q)\Big]_{11} = \Big(\mathrm{vech}\Big(\widehat{\mathcal{F}}(\Sigma^P)(q)-\mathcal{F}(\Sigma^P)(q)\Big)\Big)_1 \\
&=\sum_{|\alpha|\le m/2}K_H(\lambda_\alpha)\Big[\mathcal{F}_n(dP_1)(\alpha)\Big]\Big[\mathcal{F}_n(dP_1)(q-\alpha)\Big]-\mathcal{F}(\Sigma_{11})(q).
\end{aligned}$$

For the sake of simplicity we drop the subscript denoting asset; for example $P(t_j)$ and

n are hereafter taken to mean $P_1(t_{j,1})$ and n_1, respectively. In view of the relationship between the spectral window K and the lag window k (3.14), we have the following 'bias-variance' decomposition:

$$\begin{aligned}
\mathcal{E}_1 &= \sum_{|\alpha|\le m/2} K_H(\lambda_\alpha)\left[\sum_{j=1}^{n} e^{-\mathrm{i}\alpha t_j}\Delta P(t_j)\right]\left[\sum_{k=1}^{n} e^{-\mathrm{i}(q-\alpha)t_k}\Delta P(t_k)\right] - \int_0^{2\pi} e^{-\mathrm{i}qt} d[P,P](t) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n}\left[\sum_{|\alpha|\le\frac{m}{2}} K_H(\lambda_\alpha)\, e^{-\mathrm{i}\alpha(t_j-t_k)}\right]\Delta P(t_j)\Delta P(t_k)e^{-\mathrm{i}qt_k} - \int_0^{2\pi} e^{-\mathrm{i}qt} d[P,P](t) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} k_H(t_j-t_k)\Delta P(t_j)\Delta P(t_k)e^{-\mathrm{i}qt_k} - \int_0^{2\pi} e^{-\mathrm{i}qt}\ d[P,P](t) \\
&= \left[\sum_{j=1}^{n}\big(\Delta P(t_j)\big)^2 e^{-\mathrm{i}qt_j} - \int_0^{2\pi} e^{-\mathrm{i}qt} d[P,P](t)\right] + \left[\sum_{j\ne k} k_H(t_j-t_k)\Delta P(t_j)\Delta P(t_k)e^{-\mathrm{i}qt_k}\right] \\
&=: M_1 + M_2.
\end{aligned}$$

We first show that $\sqrt{n/H}M_1$ converges in probability to zero. Since Itô's lemma gives

$$\big[\Delta P(t_j)\big]^2 e^{-\mathrm{i}qt_j} = 2\int_{t_{j-1}}^{t_j}\{P(t)-P(t_{j-1})\}\cdot e^{-\mathrm{i}qt_j}\ dP(t) + \int_{t_{j-1}}^{t_j} e^{-\mathrm{i}qt_j} d[P,P](t),$$

we can further decompose M_1 into a martingale M_{11} and a predictable finite variation component A, where

$$\begin{aligned}
M_{11} &= 2\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\{P(t)-P(t_{j-1})\}e^{-\mathrm{i}qt_j} dP(t) = O_p(n^{-1/2}), \\
A &= \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}(e^{-\mathrm{i}tq} - e^{-\mathrm{i}t_jq})\ d[P,P](t) = O(n^{-1}).
\end{aligned}$$

This is the Euler discretization error whose distribution is given by the Theorem 5.5 of Jacod and Protter (1998). Therefore, it follows that $\sqrt{n/H}M_1 = O_p(1/\sqrt{H}) = o_p(1)$.

Now we show stable convergence of M_2. Using the off-diagonal step function $f_n(t,s;q)$ defined in Lemma 3.3:

$$f_n(t,s;q) \;=\; \sum_{j\ne k} e^{-\mathrm{i}qt_k} k_H(t_j-t_k) 1_{[t_{j-1},t_j[}(t) 1_{[t_{k-1},t_k[}(s),$$

we write

$$\begin{aligned}
M_2 &= \sum_{j\ne k} k_H(t_j-t_k)\Delta P(t_j)\Delta P(t_k)e^{-\mathrm{i}qt_k} \\
&= \int_0^{2\pi}\int_0^{2\pi} f_n(t,s;q)\ dP(s)dP(t).
\end{aligned}$$

Clearly, the expectation of M_2 is zero. Furthermore, using Lemma 3.1 we see that

$$E[M_2, M_2] \;=\; 2\int_0^{2\pi}\int_0^{2\pi} f_n^2(t, s; q)\; d[P, P](s)d[P, P](t),$$

whose limiting behaviour is given by Lemma 3.2. Consequently the distribution of $\mathcal{E}_1$ follows by Lemma 3.4, provided the conditional Lindeberg condition holds. It suffices to verify the Lyapunov condition: For some $\varepsilon > 0$, $\sum_j E(\{\chi_j^n\}^{2+\varepsilon}|\mathcal{F}_{j-1}) \to^p 0$, as we shall do here for $\varepsilon = 2$.

For later reference we note that the fourth moment of the return is given by

$$\begin{aligned}
E\left(\left[\Delta P(t_j)\right]^4\right) &= E\left(2\int_{t_{j-1}}^{t_j} \left\{P(t) - P(t_{j-1})\right\} dP(t) \;+\; \int_{t_{j-1}}^{t_j} d[P, P](t)\right)^2 \\
&= 4E\left(\int_{t_{j-1}}^{t_j} \left\{\Delta P(t_j)\right\} dP(t)\right)^2 \;+\; E\left(\int_{t_{j-1}}^{t_j} d[P, P](t)\right)^2 \\
&= 2E\int_{t_{j-1}}^{t_j}\int_{t_{j-1}}^{t} d[P, P](s)d[P, P](t) \;+\; E\left(\int_{t_{j-1}}^{t_j} d[P, P](t)\right)^2 \\
&= 3E\left(\int_{t_{j-1}}^{t_j} d[P, P]_t\right)^2
\end{aligned}$$

due to Itô's lemma and the isometry property.

Now letting

$$\chi_j^n \;:=\; \sum_{k<j}\left[\sqrt{\frac{n}{H}}\left\{\Delta P(t_j)\Delta P(t_k)k_H(t_j - t_k)\cdot\left(e^{-\mathrm{i}t_k q} + e^{-\mathrm{i}t_j q}\right)\right\}\right],$$

we can construct an upper bound of $E|\chi_j^n|^4$ for $j = n$:

$$
\begin{aligned}
E|\chi_j^n|^4 \leq\ & 2^4 \left(\frac{n}{H}\right)^2 E\left\{\sum_{h=1}^{n} \Delta P(t_j)\Delta P(t_{j-h}) \cdot k\left(\frac{t_h}{\Delta t H}\right)\right\}^4 \\
=\ & 2^4 \left(\frac{n}{H}\right)^2 \sum_{h=1}^{n} E\left([\Delta P(t_j)]^4\right) E\left([\Delta P(t_{j-h})]^4\right) \Big] k^4\left(\frac{t_h}{\Delta t H}\right) \\
& + 2^4 \cdot 6 \left(\frac{n}{H}\right)^2 \sum_{h,m=1}^{n} E\left([\Delta P(t_j)]^4\right) E\left([\Delta P(t_{j-h})]^2\right) E\left([\Delta P(t_{j-m})]^2\right) k_H^2(t_h) k_H^2(t_m) \\
=\ & 2^4 \cdot 3^2 \left(\frac{n}{H}\right)^2 \sum_{h=1}^{n} E\left(\int_{t_{j-1}}^{t_j} [P,P]'(t)dt\right)^2 \left(\int_{t_{j-h-1}}^{t_{j-h}} [P,P]'(t)dt\right)^2 k^4\left(\frac{t_h}{\Delta t H}\right) \\
& + 2^4 \cdot 6 \cdot 3 \left(\frac{n}{H}\right)^2 \left[\sum_{h,m=1}^{n} E\left(\int_{t_{j-1}}^{t_j} [P,P]'(t)dt\right)^2 E\left(\int_{t_{j-h-1}}^{t_{j-h}} [P,P]'(t)dt\right)\right. \\
& \qquad \left. \times E\left(\int_{t_{j-m-1}}^{t_{j-m}} [P,P]'(t)dt\right) k^2\left(\frac{t_h}{\Delta t H}\right) k^2\left(\frac{t_m}{\Delta t H}\right)\right] \\
& \leq 144 \cdot n^2 \times \sup_t \left\{[P,P]'(t)\right\}^4 \times \sup_{1\leq j\leq n} (\Delta {t_i}^4) \\
& \qquad \times \left[H^{-1}\frac{1}{H}\sum_{h=1}^{n} k^4\left(\frac{t_h}{\Delta t H}\right) + \frac{2}{H^2}\sum_{h=1}^{n} k^2\left(\frac{t_h}{\Delta t H}\right)\sum_{m=1}^{n} k^2\left(\frac{t_m}{\Delta t H}\right)\right] \\
=\ & n^{-2}H^{-1}C_1\left(\int_0^\infty k^4(x)dx\right) + 2n^{-2}C_2\left(\int_0^\infty k^2(x)dx\right)^2 \qquad (3.44) \\
=\ & O(n^{-2}) = o(1),
\end{aligned}
$$

provided $\int_0^\infty k^4(x)dx < \infty$, where the last line is due to Assumption 3.1. Consequently the distributional result in Lemma 3.4 follows.

We now give a result for the off-diagonal element of the estimator. When time stamps are synchronous and sample sizes are balanced, the proof is same as the univariate case. We will give a proof for the most general case; that is, when time stamps are asynchronous and sample sizes are unbalanced. We first show for the bivariate case and will extend the result to general $p \times p$ dimension. Denote the transaction time of the first asset $t_{i,1} = t_i$ and the second asset $t_{j,2} = s_j$ for the sake of simplicity. The centered estimator in (3.9) can be decomposed into, $\mathcal{E} = M_1 + M_2$, where

$$
\begin{aligned}
M_1 &= \sum_{i,j} e^{-\mathrm{i}s_j q} k_H(t_i - s_j)\Delta P_1(t_i)\Delta P_2(s_j) 1_{\{I_{i,1}\cap I_{j,2}\neq\emptyset\}} - \int_0^{2\pi} e^{-\mathrm{i}qt} d[P_1,P_2](t) \\
M_2 &= \sum_{i,j} e^{-\mathrm{i}s_j q} k_H(t_i - s_j)\Delta P_1(t_i)\Delta P_2(s_j) 1_{\{I_{i,1}\cap I_{j,2}=\emptyset\}}.
\end{aligned}
$$

We first derive the asymptotic bias. Leting $u_{ij} = t_i \wedge s_j$ and $l_{ij} = t_{i-1} \vee s_{j-1}$, we have

$$\begin{aligned} E(M_1) &= E\Bigg(\sum_{i,j} e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} dP_1(t) \int_{l_{i,j}}^{u_{i,j}} dP_2(s) 1_{\{I_{i,1} \cap I_{j,2} \neq \emptyset\}} - \int_0^{2\pi} e^{-\mathrm{i}qt} d[P_1, P_2](t) \Bigg) \\ &- E\Bigg(\sum_{i,j} e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} dP_1(t) \int_{l_{i,j}}^{u_{i,j}} dP_2(s) \{1 - k_H(t_i - s_j)\} 1_{\{I_{i,1} \cap I_{j,2} \neq \emptyset\}} \Bigg). \end{aligned}$$

By multivariate Itô calculus, conditionally on $1_{\{I_{i,1} \cap I_{j,2} = \emptyset\}}$, $E(M_1)$ can be written as the expectation of following sums

$$\sum_{i,j} e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} \{P_1(t) - P_1(l_{i,j})\} dP_2(t) + e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} \{P_2(t) - P_2(l_{i,j})\} dP_1(t) \tag{3.45}$$

$$+ \sum_{i,j} \int_{l_{i,j}}^{u_{i,j}} (e^{-\mathrm{i}s_j q} - e^{-\mathrm{i}tq}) d[P_1, P_2](t) \tag{3.46}$$

$$- \sum_{i,j} e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} dP_1(t) \int_{l_{i,j}}^{u_{i,j}} dP_2(s) \{1 - k_H(t_i - s_j)\}. \tag{3.47}$$

Using the notation of (3.4), we see the order of magnitude of the first term in (3.45) is

$$\begin{aligned} &\sum_{1 \le i \le n_1, 1 \le j \le n_2} e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} \{P_1(t) - P_1(l_{i,j})\} dP_2(t) \\ &= \sum_{l=1}^{N_T} \int_{T_{l-1}}^{T_l} \{P_1(t) - P_1(T_{l-1})\} dP_2(t) \\ &\qquad - \sum_{i,j} (1 - e^{-\mathrm{i}s_j q}) \int_{l_{i,j}}^{u_{i,j}} \{P_1(t) - P_1(l_{i,j})\} dP_2(t) \\ &= O_p(N_T^{-1/2}) + O_p(n_2^{-1} N_T^{-1/2}). \end{aligned}$$

The order of the magnitude for the second term in (3.45) is derived in a similar way. The change of discretization points to the union of the time points are without error and holds analytically. As for the second term, we discretize the deterministic function $e^{-\mathrm{i}tq}$ over the time stamp of s_j; the order of (3.46) is therefore $O_p(n_2^{-1})$, since we can replace the summation $\sum_{i,j} \int_{l_{i,j}}^{u_{i,j}}$ by $\sum_{1 \le j \le n_2} \int_{s_{j-1}}^{s_j}$. This term is zero for the integrated (co)variance estimator, $q = 0$.

Now the asymptotic bias term conditional on the volatility path is given by

$$\begin{aligned} &E\left(\sum_{i,j} e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} dP_1(t) \int_{l_{i,j}}^{u_{i,j}} dP_2(s) \{1 - k_H(t_i - s_j)\} 1_{\{I_{i,1} \cap I_{j,2} \neq \emptyset\}} \right) \\ &\simeq \sum_{i,j} e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} d[P_1, P_2](t) \left\{ -\frac{1}{2} k''(0) \left(\frac{t_i - s_j}{\widetilde{\Delta t} H} \right)^2 \right\} 1_{\{I_{i,1} \cap I_{j,2} \neq \emptyset\}} \end{aligned} \tag{3.48}$$

It is then bounded above by

$$
\begin{aligned}
(3.48) \quad \leq \quad & \left(\frac{n_1 \vee n_2}{(n_1 \wedge n_2)H}\right)^2 \frac{1}{2}\left\{\frac{n_1 \wedge n_2}{2\pi} \sup_{i,j} |t_i - s_j| 1_{\{I_{i,1}\cap I_{j,2}\neq\emptyset\}}\right\}^2 \\
& \times |k''(0)| \sum_{i,j} e^{-\mathrm{i}s_j q} \int_{l_{i,j}}^{u_{i,j}} d\,|[P_1, P_2]|\,(t) \\
= \quad & \left(\frac{n_1 \vee n_2}{(n_1 \wedge n_2)H}\right)^2 \frac{1}{2}\mathcal{A}^2 |k''(0)| \int_0^{2\pi} e^{-\mathrm{i}tq} d\,|[P_1, P_2]|\,(t). \qquad (3.49)
\end{aligned}
$$

The first approximation holds by Taylor expansion of $\{k(0) - k_H(t_i - s_j)\}$, Assumption 3.2-(3) and $k'(0) = 0$. Then the order of the stochastic bias M_1 is given by $O_p(N_T^{-1/2}) + O_p(n_2^{-1}) + O_p(\{\frac{n_1 \vee n_2}{(n_1 \wedge n_2)H}\}^2)$ for the estimator at non-zero frequency and $O_p(N_T^{-1/2}) + O_p(\{\frac{n_1 \vee n_2}{(n_1 \wedge n_2)H}\}^2)$ for the integrated (co)variance estimator. In both cases, the leading order term for the bias is the last term under the optimal bandwidth.

We next move on to M_2, and write

$$
M_2 = \int_0^{2\pi} \int_{s<t} f_n(t, s; q) dP_2(s) dP_1(t) + \int_0^{2\pi} \int_{s<t} f_n(s, t; q) dP_1(s) dP_2(t),
$$

where $f_n(t, s; q)$ is given in Lemma 3.3. Its expectation is zero, and moreover by Lemma 3.1 the expectation of the quadratic variation $E\,[M_2, M_2]$ is given by

$$
\begin{aligned}
E\,[M_2, M_2] = & \int_0^{2\pi} \int_0^{2\pi} f_n^2(t, s; q)\ d[P_2, P_2](s) d[P_1, P_1](t) && (3.50) \\
& + \int_0^{2\pi} \int_0^{2\pi} f_n(t, s; q) f_n(s, t; q)\ d[P_2, P_1](s) d[P_1, P_2](t). && (3.51)
\end{aligned}
$$

Now using similar arguments of the proof of Lemma 3.2, we can show that

$$
\begin{aligned}
& \frac{n_1 \wedge n_2}{H} \int_0^{2\pi} \int_0^{2\pi} f_n^2(t, s; q) d[P_1, P_1](t) d[P_2, P_2](s) \\
& \qquad \longrightarrow \|k\|^2 \int_0^{2\pi} e^{-\mathrm{i}2tq} [P_1, P_1]'(t) [P_2, P_2]'(t) d\mathcal{Q}_{1122}(t) && (3.52) \\
& \frac{n_1 \wedge n_2}{H} \int_0^{2\pi} \int_0^{2\pi} f_n(t, s; q) f_n(s, t; q) d[P_1, P_2](t) d[P_2, P_1](s) \\
& \qquad \longrightarrow \|k\|^2 \int_0^{2\pi} e^{-\mathrm{i}2tq} ([P_1, P_2]')^2(t) d\mathcal{Q}_{1212}(t). && (3.53)
\end{aligned}
$$

Hence the proof of Proposition 3.2 is complete in view of (3.49), (3.52) and (3.53).

Now similar arguments as in the univariate case yields $\sup_i E|\chi_i^n|^4 = O((n_1 \wedge n_2)^{-2})$, where

$$
\chi_i^n = \sum_{j: s_j < t_i} \sqrt{m} \Delta P_1(t_i) \Delta P_2(s_j) k_H(t_i - s_j) e^{-\mathrm{i}s_j q} 1_{\{I_{i,1}\cap I_{j,2}=\emptyset\}};
$$

so stable convergence can be established using Lemma 3.4 and Proposition 3.1. Higer-dimensional extension of the asymptotic results involves the use of the Cramér-Wold device; it is sufficient to show that any linear combination of the elements of the matrix estimator converges to the corresponding univariate Gaussian random variable. Denote by $\mathcal{R}(q) := \widehat{\mathcal{F}}(\Sigma)(q) - \mathcal{F}(\Sigma)(q)$ and consider the linear combinations $a^\top \mathcal{R}(q) b$ and $c^\top \mathcal{R}(q) d$, where a, b, c, d are some arbitrary constant vectors of conformable dimension. Note that $a^\top \mathcal{R}(q) c b^\top \mathcal{R}(q) d = \text{vech}(ab^\top)^\top (\mathcal{R}(q) \otimes \mathcal{R}(q)) \text{vech}(dc^\top)$, and that its expectation depends on $E\{\mathcal{R}(q) \otimes \mathcal{R}(q)\}$. So the same arguments as above using lemmas in Section 3.6.1 complete the proof. ■

3.6.3 Proof of Theorem 3.2

By the triangle inequality we have

$$\begin{aligned}
&\left\| \Sigma(t) - \frac{1}{2\pi} \sum_{|q| \le m/2} K_H(\lambda_q) \exp(\mathrm{i}qt) \widehat{\mathcal{F}}(\Sigma)(q) \right\|_2 \\
&\quad \le \left\| \Sigma(t) - \frac{1}{2\pi} \sum_{|q| \le m/2} K_H(\lambda_q) \exp(\mathrm{i}qt) \mathcal{F}(\Sigma)(q) \right\|_2 \\
&\quad + \left(\frac{1}{2\pi} \right)^2 \sum_{|q|, |\ell| \le m/2} K_H(\lambda_q) K_H(\lambda_\ell) \exp\big(\mathrm{i}(q+\ell)t\big) \left\| \mathcal{F}(\Sigma)(q) - \widehat{\mathcal{F}}(\Sigma)(q) \right\|_2 .
\end{aligned}$$

where $m = n/H$. Now Theorem 1 implies that (3.9) converges in probability to $\mathcal{F}(\Sigma)(q)$ for each q. So if we assume the modulus of continuity of $\Sigma(t)$ is available and is given by (3.23) then there exists a sequence $\delta(n) \to 0$ such that

$$\sup_{\delta(n) \le t \le 2\pi - \delta(n)} \left\| \Sigma(t) - \sum_{|q| \le m/2} K_H(\lambda_q) \exp(\mathrm{i}qt) \widehat{\mathcal{F}}(\Sigma)(q) \right\|_2 \le \mathcal{C} \left(\frac{4}{m} \right),$$

as required. ■

Table 3.1: Realized Covariance

		Realized Covariance 5 min aligned					
		BIAS			rMSE		
Sampling	NoiseSignal	(1,1)	(2,2)	**(1,2)**	(1,1)	(2,2)	**(1,2)**
Equal	0	(0.01)	0.00	**(0.00)**	0.17	0.17	**0.16**
	0.001	0.02	0.03	**0.01**	0.17	0.18	**0.16**
	0.01	0.24	0.26	**0.13**	0.31	0.34	**0.23**
(3/2,30)	0	(0.00)	0.01	**(0.08)**	0.17	0.17	**0.17**
	0.001	0.02	0.03	**(0.08)**	0.17	0.18	**0.17**
	0.01	0.25	0.26	**(0.07)**	0.32	0.33	**0.18**
(3/2,2)	0	(0.00)	0.01	**(0.00)**	0.17	0.17	**0.16**
	0.001	0.03	0.03	**0.01**	0.18	0.18	**0.16**
	0.01	0.27	0.26	**0.09**	0.34	0.35	**0.21**
(20,30)	0	(0.00)	0.01	**(0.07)**	0.17	0.17	**0.17**
	0.001	0.02	0.02	**(0.07)**	0.17	0.17	**0.18**
	0.01	0.24	0.22	**(0.04)**	0.31	0.30	**0.19**

		Realized Covariance Refresh Time aligned					
		BIAS			rMSE		
Sampling	NoiseSignal	(1,1)	(2,2)	**(1,2)**	(1,1)	(2,2)	**(1,2)**
Equal	0	(0.00)	0.00	**(0.00)**	0.01	0.01	**0.01**
	0.001	4.52	4.54	**2.10**	4.52	4.54	**2.10**
	0.01	45.24	45.41	**21.07**	45.24	45.41	**21.07**
(3/2,30)	0	0.01	0.01	**(0.02)**	0.08	0.08	**0.08**
	0.001	0.23	0.23	**0.07**	0.25	0.24	**0.10**
	0.01	2.24	2.25	**0.92**	2.25	2.26	**0.93**
(3/2,2)	0	(0.00)	0.00	**(0.17)**	0.02	0.02	**0.17**
	0.001	1.78	1.80	**0.53**	1.78	1.80	**0.53**
	0.01	17.83	17.98	**6.77**	17.84	17.99	**6.78**
(20,30)	0	0.01	0.00	**(0.27)**	0.09	0.07	**0.28**
	0.001	0.16	0.15	**(0.24)**	0.18	0.17	**0.25**
	0.01	1.55	1.55	**0.08**	1.56	1.56	**0.15**

Table 3.2: 2 dimensional covariation matrix - continuous SV ($\cdot$/100)

		Realized Cov Refresh Time			Realized Cov 5 min			HY	Realized Kernel Refresh Time			**FRK average H**			**FRK opt H**		
NRS		(1,1)	(1,2)	(2,2)	(1,1)	(1,2)	(2,2)	(1,2)	(1,1)	(1,2)	(2,2)	(1,1)	(1,2)	(2,2)	(1,1)	(1,2)	(2,2)
	Sampling: (3/2,30) Balanced																
0	bias	1.0	(1.9)	0.8	**(0.1)**	(7.6)	0.9	1.0	1.0	(1.3)	**0.8**	0.9	**(0.6)**	1.4	0.9	**(0.6)**	1.4
	rmse	**8.0**	7.6	8.1	17.1	17.1	16.8	7.6	8.2	**7.5**	**8.0**	20.9	19.6	20.8	20.9	19.6	20.8
0.001	bias	23.2	7.3	22.9	2.2	(7.7)	3.3	10.3	1.6	0.9	1.6	1.0	0.8	1.7	**0.9**	**(0.1)**	**0.1**
	rmse	24.7	**10.6**	24.6	17.2	17.2	17.6	12.9	**16.7**	15.2	**15.4**	30.2	28.9	32.3	38.2	35.5	41.2
0.01	bias	225	92.3	225	24.5	(6.6)	25.5	96.4	3.5	2.3	4.0	**0.6**	**(1.2)**	**(1.5)**	**0.6**	**(1.2)**	**(1.5)**
	rmse	226	93.6	226	31.6	17.9	33.0	97.6	25.3	23.7	25.5	43.8	40.1	46.7	43.8	40.1	46.7
	Sampling: (3/2,30) Unbalanced																
0	bias	0.1	(2.6)	**0.5**	(0.9)	(8.2)	0.9	0.4	**0.0**	(1.9)	0.6	1.0	(3.1)	1.4	1.1	**(0.1)**	1.3
	rmse	**6.7**	6.9	**7.0**	17.4	18.0	17.8	**6.6**	7.0	7.0	7.3	15.9	14.8	15.9	22.2	21.0	23.1
0.001	bias	23.5	6.6	21.3	1.4	(8.0)	3.8	9.5	2.0	**1.8**	3.1	1.5	(12.5)	5.0	**1.0**	(4.2)	**2.9**
	rmse	24.7	**9.6**	22.6	17.9	18.5	18.7	11.7	15.7	14.6	15.6	**10.5**	15.5	**12.2**	15.2	14.4	15.3
0.01	bias	239	92.5	210	23.5	(6.9)	27.3	95.7	2.6	2.3	**4.6**	1.9	(1.9)	11.6	**1.6**	**0.3**	4.9
	rmse	240	93.9	211	31.7	20.3	35.2	96.8	24.0	23.3	25.9	**18.1**	**16.7**	**21.7**	23.9	22.9	25.8
	Sampling: (20,30) Unbalanced																
0	bias	**0.1**	(27.1)	0.7	0.6	(6.7)	1.3	0.8	0.2	(23.7)	**0.7**	0.9	**(0.4)**	1.8	0.9	**(0.4)**	1.8
	rmse	**8.3**	28.0	**8.2**	17.2	17.5	17.8	**9.3**	8.4	24.8	8.2	20.2	19.1	20.9	20.2	19.1	20.9
0.001	bias	16.8	(23.1)	16.2	3.1	(6.1)	3.7	4.9	2.0	0.1	2.8	1.1	**0.1**	1.9	**0.8**	0.4	**1.5**
	rmse	19.0	24.4	18.6	18.2	18.0	18.8	**11.0**	**16.3**	15.0	**16.2**	23.3	22.1	24.2	26.9	25.9	28.7
0.01	bias	162	10.4	155	26.1	(1.7)	26.0	41.1	3.8	2.2	4.7	1.7	0.6	1.9	**1.0**	**(0.5)**	**0.2**
	rmse	164	**16.5**	156	34.0	20.4	34.3	43.6	**24.3**	23.4	26.7	32.2	30.8	34.5	35.8	34.1	39.4

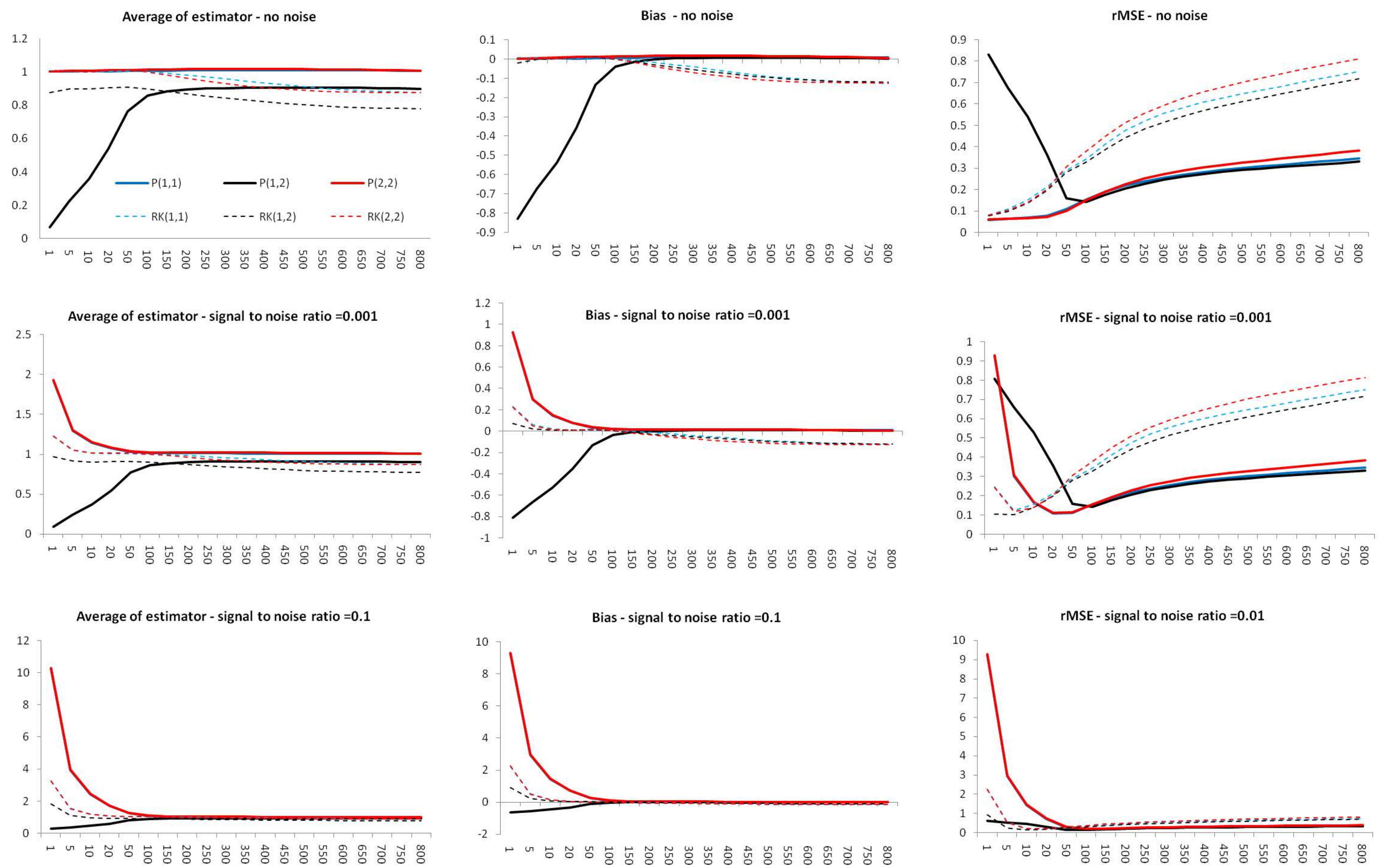
Table 3.3: 2 dimensional covariation matrix - jump diffusion SV ($\cdot$/100)

		Realized Cov Refresh Time			Realized Cov 5 min			HY	Realized Kernel Refresh Time			**FRK average H**			**FRK opt H**		
NRS		(1,1)	(1,2)	(2,2)	(1,1)	(1,2)	(2,2)	(1,2)	(1,1)	(1,2)	(2,2)	(1,1)	(1,2)	(2,2)	(1,1)	(1,2)	(2,2)
	Sampling: (3/2,30) Balanced																
0	bias	(0.3)	(2.0)	**(18.7)**	(2.2)	(6.0)	(20.4)	**(0.3)**	(0.2)	(1.5)	(18.6)	**(0.1)**	(0.7)	(18.8)	**(0.1)**	(0.7)	(18.8)
	rmse	**6.7**	**18.1**	20.3	17.2	21.9	25.4	18.6	7.1	18.3	**20.3**	18.0	23.4	24.6	18.0	23.4	24.6
0.001	bias	23.3	8.2	**4.9**	0.5	(5.6)	(17.2)	9.8	1.5	1.3	(16.9)	**0.0**	**0.1**	(19.3)	**0.0**	**0.1**	(19.3)
	rmse	24.7	**19.8**	**9.9**	16.8	22.0	23.0	21.1	**13.4**	22.0	21.8	29.1	29.2	31.9	29.1	29.2	31.9
0.01	bias	233	96.7	211	24.2	(3.6)	**7.0**	98.8	2.6	2.1	(15.8)	1.4	1.3	(18.0)	**10.2**	**(0.7)**	**(10.0)**
	rmse	234	99.6	212	30.8	22.6	19.9	101.8	19.8	24.5	25.3	42.6	38.9	42.8	**17.2**	**21.2**	**17.1**
	Sampling: (3/2,30) Unbalanced																
0	bias	(0.3)	(1.9)	**(18.8)**	(1.6)	(6.6)	(21.2)	**(0.1)**	(0.2)	(1.6)	(18.8)	0.2	(2.5)	(19.1)	**0.0**	(0.5)	(18.9)
	rmse	7.1	**17.7**	**20.4**	16.9	22.0	26.1	18.2	**7.1**	17.9	20.5	13.7	21.1	23.0	20.1	24.4	25.3
0.001	bias	24.0	8.0	**2.6**	0.8	(6.6)	(19.0)	10.0	0.7	**0.5**	(17.5)	1.5	(8.3)	(14.6)	**0.7**	(3.2)	(17.5)
	rmse	25.5	19.7	**9.3**	17.6	22.4	24.7	21.1	14.0	20.7	21.5	9.2	**19.1**	18.0	**12.8**	20.5	21.6
0.01	bias	243	94.1	192	23.1	(6.1)	3.4	97.8	**1.4**	**1.0**	(15.7)	1.8	(1.6)	(8.6)	3.8	(3.3)	**2.6**
	rmse	244	97.1	193	32.1	24.2	20.8	100.7	22.1	24.7	23.8	15.2	21.6	17.3	**12.6**	**19.6**	**14.8**
	Sampling: (20,30) Unbalanced																
0	bias	(0.1)	(16.7)	**(18.0)**	(1.4)	(5.4)	(20.1)	**(0.0)**	0.1	(14.7)	(18.0)	**(0.0)**	(0.9)	(18.9)	(0.2)	(0.1)	(18.9)
	rmse	**6.8**	21.5	**19.5**	16.8	21.7	26.3	**19.2**	6.8	20.4	19.6	17.6	23.3	24.2	24.2	27.0	28.3
0.001	bias	15.3	(13.2)	**(2.7)**	1.4	(4.7)	(17.6)	3.7	1.6	(0.4)	(17.9)	0.1	(0.4)	(18.5)	**0.1**	**(0.0)**	(18.8)
	rmse	17.1	**19.1**	**8.9**	17.1	21.6	24.7	19.6	**14.5**	21.2	21.7	19.7	24.5	25.1	23.9	26.7	28.0
0.01	bias	159	20.6	136	24.9	**0.1**	**4.9**	40.1	3.0	1.2	(15.9)	0.8	0.9	(17.8)	6.2	**0.7**	**(10.5)**
	rmse	**160**	26.8	137	32.2	**22.3**	20.9	45.5	22.7	25.8	24.1	30.2	30.6	32.6	17.9	**22.6**	**17.7**

Table 3.4: Scalar function of 10 dimensional covariation matrix

		max (eigenvalue)		portfolio	
Noise to Signal Ratio=0		Bias	rMSE	Bias	rMSE
RV_refresh		(2.34)	2.75	1.76	2.74
RV_fixed		**(0.85)**	3.18	**0.14**	4.09
Realized Kernel		(2.21)	2.65	1.66	2.67
Fourier RK		(1.18)	**2.17**	0.26	**2.51**
Noise to Signal Ratio=0.001					
RV_refresh		7.22	7.50	27.20	27.47
RV_fixed		0.40	3.28	4.14	6.31
Realized Kernel		0.38	3.00	1.67	4.16
Fourier RK	minH	**(0.28)**	**1.95**	3.88	4.87
	avgH	(0.47)	2.64	0.73	**3.67**
	maxH	(0.52)	3.13	**(0.20)**	3.99
Noise to Signal Ratio=0.01					
RV_refresh		127.24	127.81	256.02	257.05
RV_fixed		15.42	16.54	40.37	41.99
Realized Kernel		1.29	4.82	3.23	6.86
Fourier RK	minH	1.22	**3.27**	6.88	8.55
	avgH	0.13	4.02	1.76	**5.58**
	maxH	**(0.03)**	4.92	**0.67**	5.96

Figure 3.6: Simulation Result : Balanced, Sampled at $= \{3/2, 30\}$. [a]



[a] Comparison of 2×2 covariance matrix and its scalar function estimated by Realized Kernel and Fourier Realized Kernel for range of bandwidths. Sample size of two assets are unbalanced and time stamps are poisson sampled at rate $= \{3/2, 30\}$.

Figure 3.7: Simulation Result : Unbalanced, Sampled at $= \{3/2, 30\}$

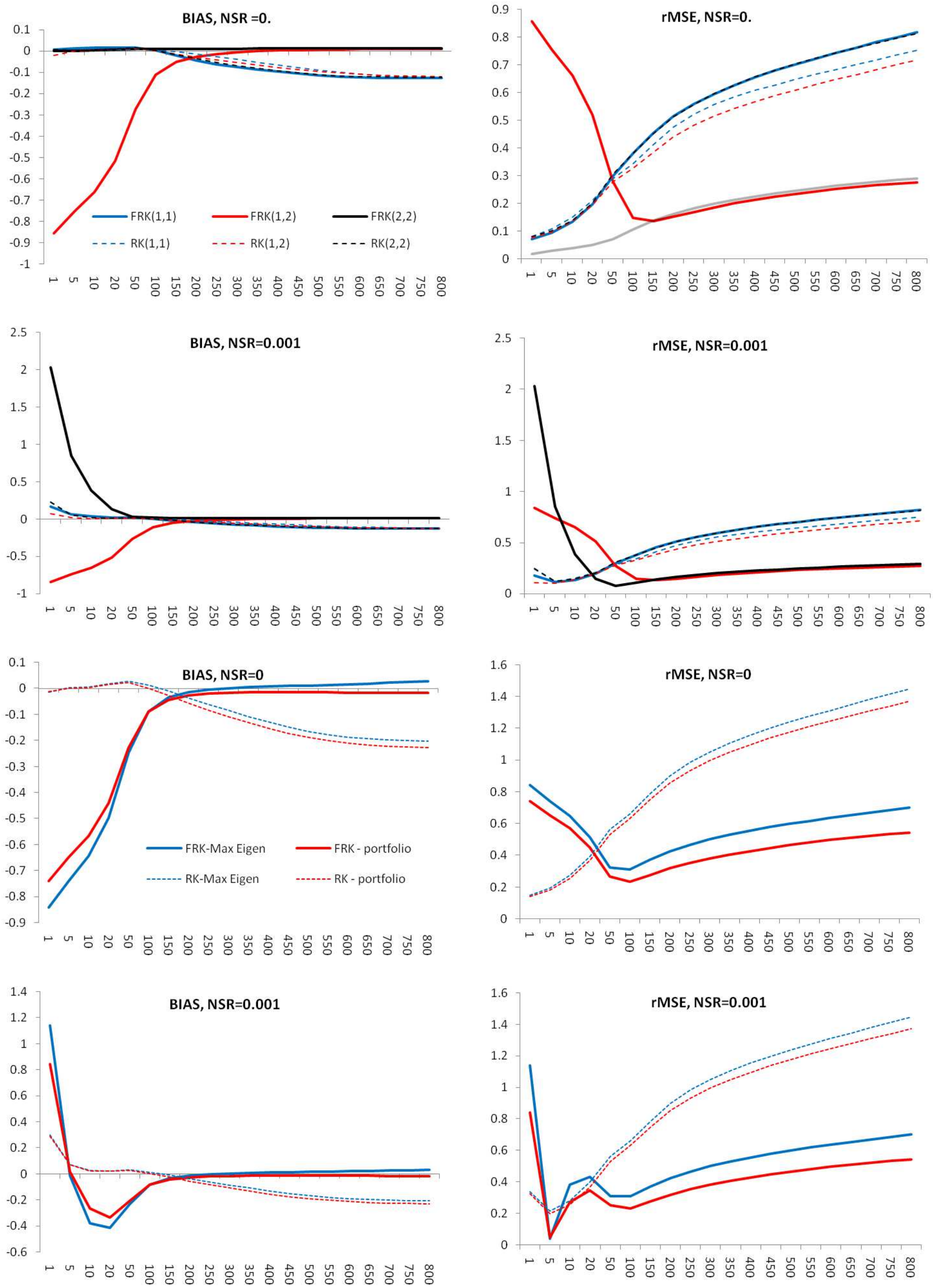


Figure 3.8: Simulation Result : Unbalanced, Sampled at $= \{20, 30\}$

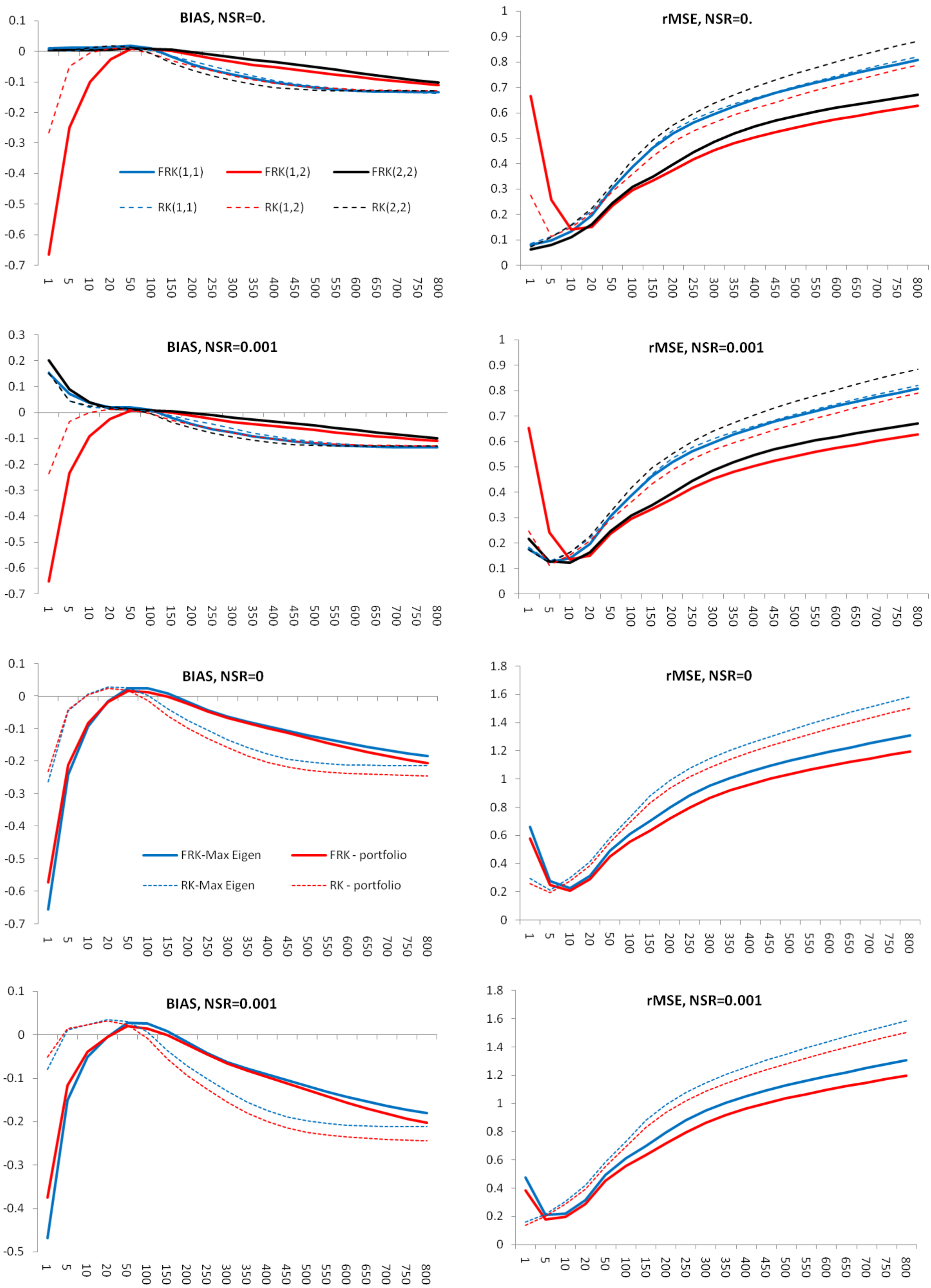
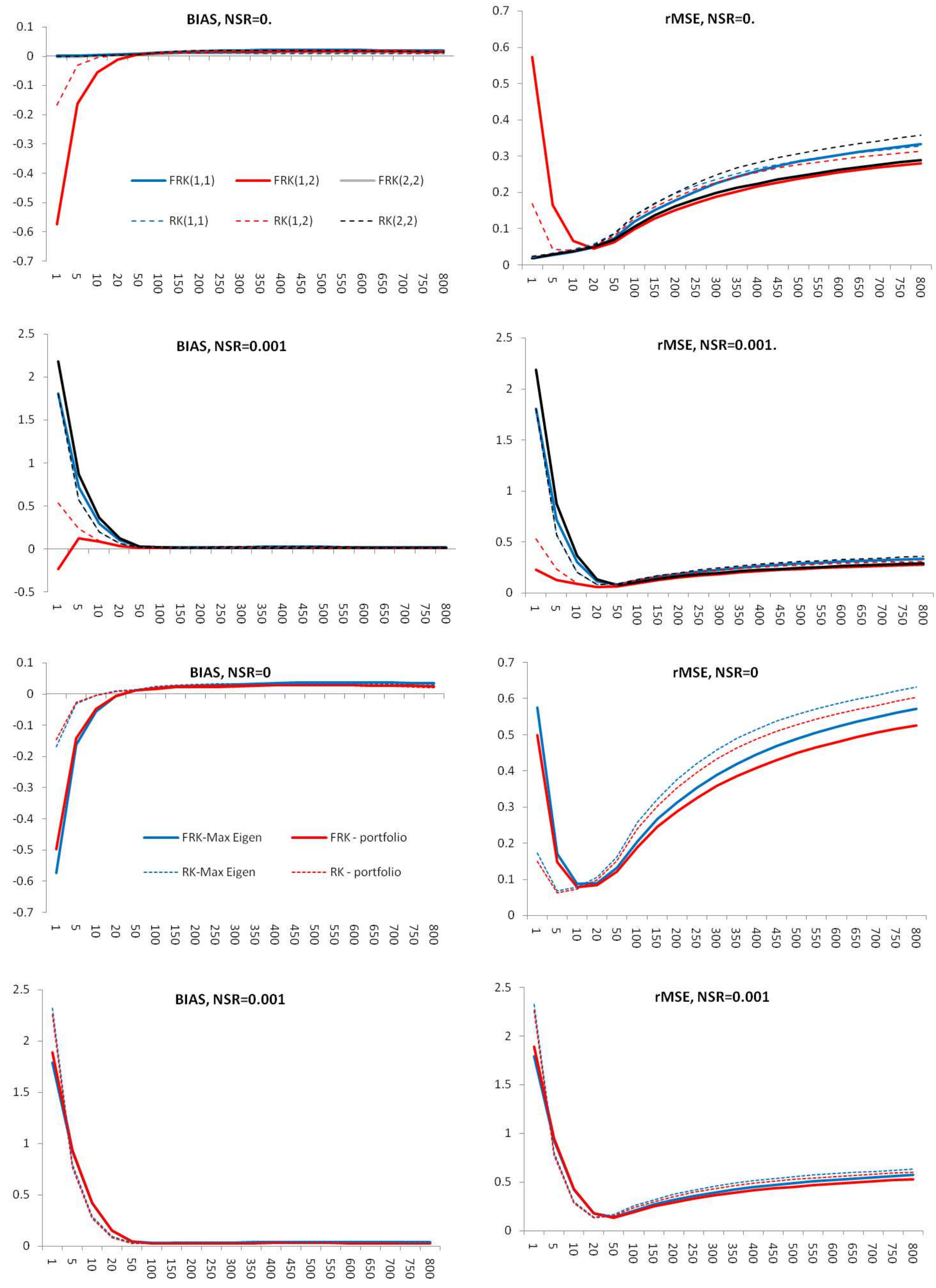


Figure 3.9: Simulation Result : Unbalanced, Sampled at $= \{3/2, 2\}$



Bibliography

[1] Abel, A. B. (1988): "Stock prices under time-varying dividend risk : An exact solution in an infinite-horizon general equilibrium model," Journal of Monetary Economics, 22(3), 375–393.

[2] Aït-Sahalia, Y. and Jacod, J. (2014). *High-frequency Financial Econometrics.* New Jersey: Princeton University Press.

[3] Aït-Sahalia, Y., Fan, J. and Xiu, D. (2010). *High-frequency covariance estimates with noisy and asynchronous financial data.* Journal of the American Statistical Association, 105(492), 1504-1517.

[4] Aït-Sahalia, Y., Mykland, P. A. and Zhang, L. (2011). *Ultra high frequency volatility estimation with dependent microstructure noise.* Journal of Econometrics, 160(1), 160-175.

[5] Aldous, D. J. and Eagleson, G. K. (1978). *On mixing and stability of limit theorems.* Annals of Probability, 6(2), 325-331.

[6] Anderson, R. M. (2011). *Time-varying risk premia.* Journal of Mathematical Economics, 47(3), 253-259.

[7] Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2001). *The distribution of realized exchange rate volatility.* Journal of the American Statistical Association, 96(453), 42-55.

[8] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* New Jersey: Wiley & Sons.

[9] Andrews, D. W. K. (1984). *Non-Strong Mixing Autoregressive Processes.* Journal of Applied Probability, 21(4), 930-934.

[10] Andrews, D. W. K. (1991). *Heteroskedasticity and autocorrelation consistent covariance matrix estimation.* Econometrica, 59(3), 817-858.

[11] Andrews, D. W. K. (1995). *Nonparametric kernel estimation for semiparametric models.* Econometric Theory, 11(3), 560-586.

[12] Aneiros, G., Bongiorno, E. G., Cao, R. and Vieu, P. (eds.). (2017). *Functional Statistics and Related Fields*. New York: Springer.

[13] Antell, J. and Vaihekoski, M. (2016). *Countercyclical and Time-Varying Risk Aversion and the Equity Premium*. SSRN Working paper available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2753537.

[14] Azais, J. M. and Fort, J. C. (2013). *Remark on the finite-dimensional character of certain results of functional statistics*. Comptes Rendus Mathematique, 351(3), 139-141.

[15] Bachelier, L. (1900). *Théorie de la spéculation*. Annales Scientifiques de l'École Normale Supérieure Sér. 3(17), 21-86.

[16] Backus, K. and Gregory, A, W. (1993). *Theoretical Relations Between Risk Premiums and Conditional variances*. Journal of Business and Economic Statistics, 11(2), 177-185.

[17] Bailey, N., Kapetanios, G. and Pesaran, H. (2012). *Exponent of Cross-sectional Dependence: Estimation and Inference* (January 2012). Working paper available at http://www.econ.cam.ac.uk/research/repec/cam/pdf/cwpe1206.pdf

[18] Barndorff-Nielsen, O. E. and Shephard, N. (2004). *Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics*. Econometrica, 72(3), 885-925.

[19] Bali, T. G. and Peng, L. (2006). *Is there a Risk-Return Trade-Off? Evidence from High-frequency data*. Journal of Applied Econometrics, 21, 1169-1198.

[20] Bansal, R. and Yaron, A. (2004): *Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles*. Journal of Finance, 59(4), 1481-1509.

[21] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008). *Designing realized kernels to measure the ex post variation of equity prices in the presence of noise*. Econometrica, 76(6), 1481-1536.

[22] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2011). *Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading*. Journal of Econometrics, 162(2), 149-169.

[23] Barndorff-Nielsen, O. E. and Shephard, N. (2002). *Econometric analysis of realized volatility and its use in estimating stochastic volatility models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(2), 253-280.

[24] Bibinger, M., Hautsch, N., Malec, P. and Reiss, M. (2014). *Estimating the quadratic covariation matrix from noisy observations: local method of moments and efficiency.* Annals of Statistics, 42(4), 1312-1346.

[25] Bierens, H. J. (1983). *Uniform consistency of kernel estimators of a regression function under generalized conditions.* Journal of the American Statistical Association, 78(383), 699-707.

[26] Bierens, H. J. (1987). *Kernel estimators of regression functions.* In Bewley, T. F. (ed.) *Advances in econometrics: Fifth world congress, 99-144, Vol. 1*, Cambridge: Cambridge University Press.

[27] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987). *Regular variation.* Cambridge: Cambridge University Press.

[28] Billingsley, P. (1968). *Convergence of Probability Measures.* New York: John Wiley.

[29] Blanchard, O. J. and Watson, M. W. (1982). *Bubbles, rational expectations and financial markets.* NBER Working Paper no. 945 available at http://www.nber.org/papers/w0945.

[30] Bliss, R. R. and Panigirtzoglou, N. (2004). *Option-implied risk aversion estimates.* Journal of Finance, 59(1), 407-446.

[31] Borovkov, A. A. and Ruzankin, P. S. (2008). *On small deviations of series of weighted random variables.* Journal of Theoretical Probability, 21(3), 628-649.

[32] Bollerslev, T. (1996). *Modeling and pricing long memory in stock market volatility.* Journal of Econometrics, 73(1), 151-184.

[33] Bollerslev, T., Gibson, M. and Zhou, H. *Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities.* Journal of Econometrics, 160(1), 235-245.

[34] Bosq, B. (1996). *Nonparametric statistics for stochastic processes: estimation and prediction.* New York: Springer-Verlag.

[35] Boudoukh, J., Richardson, M. P. and Whitelaw, R. F. (1994). *A tale of three schools: Insights on autocorrelations of short-horizon stock returns.* Review of Financial Studies, 7(3), 539-573.

[36] Boudoukh, J., Richardson, M. P. and Whitelaw, R. F. (1997). *Nonlinearities in the Relation between the Equity Risk Premium and the Term Structure.* Management Science, 43(3), 371-385.

[37] Bradley, R. C. (2005). *Basic properties of strong mixing conditions. A survey and some open questions*. Probability Surveys, 2(2), 107-144.

[38] Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods*. New York: Springer-Verlag.

[39] Campbell, J. Y. and Hentschel, L. (1992). *No News is Good News: An Asymmetric Model of Changing Volatility in Stock Returns*. Journal of Financial Economics, 31(3), 281-318.

[40] Campbell, J. Y. (1987). *Stock Returns and the Term Structure*. Journal of Financial Economics, 18(2), 373-399.

[41] Campbell, J. Y. and Cochrane, J. H. (1999). *By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior*. Journal of Political Economy, 107(2), 205-251.

[42] Campbell, J. Y., Lo, A. W. and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press.

[43] Castura, J., Litzenberger, R., Gorelick, R. and Dwivedi, Y. (2010). *Market efficiency and microstructure evolution in US equity markets: A high-frequency perspective*. Working paper available at http://www.sec.gov/comments/s7-02-10/s70210-364.pdf.

[44] Charles, A. and Darné, O. (2009). *Variance-Ratio Tests of Random Walk: An Overview*. Journal of Economic Surveys, 23(3), 503-527.

[45] Chaté, H. and Courbage, M. (1997). *Special issue on lattice dynamics*. Physica D, 103, 1-611.

[46] Chen, J., Li, D., Linton, O. B. and Lu, Z. (2017). *Semiparametric Ultra-High Dimensional Model Averaging of Nonlinear Dynamic Time Series*. Journal of the American Statistical Association, Forthcoming.

[47] Chen, R. and Tsay, R. S. (1993). *Nonlinear additive ARX models*. Journal of the American Statistical Association, 88(423), 955-967.

[48] Chen, W. W. and Deo, R. S. (2006). *The variance ratio statistic at large horizons*. Econometric Theory, 22(2), 206-234.

[49] Chen, X. and Christensen, T. M. (2015). *Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions*. Journal of Econometrics 188, 447-465.

[50] Chitturi, R. V. (1974). *Distribution of residual autocorrelations in multiple autoregressive schemes.* Journal of the American Statistical Association, 69(348), 928-934.

[51] Cho, J. S. and White, H. (2014). *Testing the Equality of Two Positive-Definite Matrices with Application to Information Matrix Testing.* Advances in Econometrics: Essays in Honor of Peter C. B. Phillips. Vol. 33. Eds. Yoosoon Chang, Thomas B. Fomby, and Joon Y. Park. West Yorkshire, UK: Emerald Group Publishing Limited, forthcoming.

[52] Christensen, B. J., Dahl, C. M. and Iglesias, E. M. (2012). *Semiparametric inference in a GARCH-in-mean model.* Journal of Econometrics, 167(2), 458-472.

[53] Christensen, K., Kinnebrock, S. and Podolskij, M. (2010). *Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data.* Journal of Econometrics, 159(1), 116-133.

[54] Chordia, T., Roll, R. and Subrahmanyam, A. (2011). *Recent trends in trading activity and market quality.* Journal of Financial Economics, 101(2), 243-263.

[55] Cochrane, J. H. (2001). Asset Pricing. Princeton, NJ: Princeton University Press.

[56] Cohn, A., Engelmann, J., Fehr, E. and Maréchal, M. (2015). *Evidence for Countercyclical Risk Aversion: An Experiment with Financial Professionals.* American Economic Review, 105(2), 860-885.

[57] Comte, F. and Renault, E. (1998). *Long memory in continuous-time stochastic volatility models.* Mathematical Finance, 8(4), 291-323.

[58] Conrad, C. and Mammen, E. (2008). *Nonparametric regression on latent covariates with an application to semiparametric GARCH-in-Mean models.* Discussion Paper No. 473, University of Heidelberg, Department of Economics.

[59] Cont, R. (2000). *Empirical properties of asset returns: stylized facts and statistical issues.* Quantitative Finance, 1, 223-236.

[60] Curato, I. V., Mancino, M. E. and Sanfelici, S. (2014). *High Frequency Volatility of Volatility Estimation Free from Spot Volatility Estimates.* SSRN Working Paper available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2437989

[61] Dahlhaus, R. (1997). *Fitting time series models to nonstationary processes.* Annals of Statistics, 25(1), 1-37.

[62] Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians.* Oxford: Oxford University Press.

[63] Davydov, Y. A. (1968). *Convergence of distributions generated by stationary stochastic processes.* Theory of Probability and Its Applications, 13(4), 691-696.

[64] Delsol, L. (2009). *Advances on asymptotic normality in non-parametric functional time series analysis.* Statistics, 43(1), 13-33.

[65] Deo, R. and Richardson, M. (2003). *On the asymptotic power of the variance ratio test.* Econometric Theory, 19(2), 231-239.

[66] Devroye, L. P. (1978). *The uniform convergence of the Nadaraya-Watson regression function estimate.* Canadian Journal of Statistics, 6(2), 179-191.

[67] Devroye, L. (1981). *On the almost everywhere convergence of nonparametric regression function estimates.* Annals of Statistics, 9(6), 1310-1319.

[68] Doukhan, P. (1994). *Mixing.* New York: Springer.

[69] Doukhan, P. and Wintenberger, O. (2008). *Weakly dependent chains with infinite memory.* Stochastic Processes and their Applications, 118(11), 1997-2013.

[70] Duflo, M. (1997). *Random iterative models.* Berlin: Springer-Verlag.

[71] Dufour, J.-M. and Roy, R. (1985). *Some robust exact results on sample autocorrelations and test of randomness.* Journal of Econometrics, 29(3), 257-273.

[72] Dunker, T., Lifshits, M. A. and Linde, W. (1998). *Small deviation probabilities of sums of independent random variables.* In: Eberlein, E. (ed.), High dimensional probability, volume 43 of Progress in Probability., Birkhauser, Basel, 59-74.

[73] Eaton, M. L. and Tyler, D. E. (1991). *On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix.* Annals of Statistics, 19(1), 260-271.

[74] Engle, R. F. (1984). *Wald, Likelihood ratio, and Lagrange multiplier tests in Econometrics.* in Handbook of Econometrics, Vol. II, ed. Z. Amsterdam: North Holland.

[75] Escanciano, J. C. and Lobato, I. N. (2009). *Testing the martingale hypothesis.* in Palgrave Handbook of Econometrics. Palgrave Macmillan UK.

[76] Escanciano, J.C., Pardo-Fernández, J.C. and Van Keilegom, I. (2017). *Semiparametric estimation of risk return relationships.* Journal of Business and Economic Statistics 35(1), 40-52.

[77] Escanciano, J. C. and Velasco, C. (2006). *Generalized spectral tests for the martingale difference hypothesis.* Journal of Econometrics, 134(1), 151-185.

[78] Fan, J. and Masry, E. (1992). *Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes*. Journal of Multivariate Analysis, 43(2), 237-271.

[79] Fama, E. (1970). *Effient Capital Markets: A Review of Theory and Empirical Work*. Journal of Finance 25(2), 383-417.

[80] Fama, E. (2013). *Two pillars of asset pricing*. Nobel prize lecture.

[81] Fan, J. (1990). *A remedy to regression estimators and nonparametric minimax efficiency*. Technical Report 161, Department of Statistics, University of North Carolina at Chapel Hill.

[82] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

[83] Faust, J. (1992). *When are Variance Ratio Tests for Serial Dependence Optimal?*. Econometrica, 60(5), 1215-1226.

[84] Feller, W. (1971). *Introduction to Probability Theory and Its Applications*, Vol. 2. New York: Wiley.

[85] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2010). *Rate of uniform consistency for nonparametric estimates with functional variables*. Journal of Statistical Planning and Inference, 140(2), 335-352.

[86] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2011). *Kernel regression with functional response*. Electronic Journal of Statistics, 5, 159-171.

[87] Ferraty, F. and Romain, Y. (2010). *The Oxford Handbook of Functional Data Analysis*. New York: Oxford University Press.

[88] Ferraty, F. and Vieu, P. (2002). *The functional nonparametric model and application to spectrometric data*. Computational Statistics, 17(4), 545-564.

[89] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: Theory and Practice*. New York: Springer.

[90] Foster, D. and Nelson, D. (1996). *Continuous record asymptotics for rolling sample variance estimators*. Econometrica, 64(1), 139-174.

[91] Flood, R. P. and Hodrick, R. J. (1990). *On testing for speculative bubbles*. Journal of Economic Perspectives, 4(2), 85-101.

[92] Francq, C., Roy, R. and Zakoian, J,-M. (2005). *Diagnostic Checking in ARMA Models with Uncorrelated Errors.* Journal of the Americal Statistical Association, 100(470), 532-544.

[93] French, K. R., Schwert, G. W. and Stambaugh, R. F. (1987) *Expected Stock Returns and Volatility.* Journal of Financial Economics, 19(1), 3-29.

[94] Fuk, D. K. and Nagaev, S. V. (1971). *Probability inequalities for sums of independent random variables.* Theory of Probability and its applications, 16(4), 643-660.

[95] Gao, F., Hannig, J. and Torcaso, F. (2003). *Comparison Theorems for Small Deviations of Random Series.* Electronic Journal of Probability, 8(21), 1-17.

[96] Gennotte, G. and Marsh, T. A. (1993). *Variations in economic uncertainty and risk premiums on capital assets.* European Economic Review, 37(5), 1021-1041.

[97] Geenens, G. (2011). *Curse of dimensionality and related issues in nonparametric functional regression.* Statistics Surveys, 5, 30-43.

[98] Ghysels, E., Santa-Clara, P. and Valkanov, R. (2005). *There is a Risk-Return Trade-Off After All.* Journal of Financial Economics, 76(3), 509-548.

[99] Giraitis, L., Leipus, R. and Surgailis, D. (2008). *ARCH*(∞) *and long memory properties.* In: Andersen, T. G., Davis, R. A., Kreiß, J., Mikosch, T. (Eds.), Handbook of Financial Time Series. Berlin: Springer-Verlag.

[100] Glosten, L., Jagannathan, R. and Runkle, D. E. (1993). *On The Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks.* Journal of Finance, 48(5), 1779-1801.

[101] Götze, F. and Hipp, C. (1994). *Asymptotic distribution of statistics in time series.* Annals of Statistics, 22(4), 2062-2088.

[102] Greblicki, W. and Krzyzak, A. (1980). *Asymptotic properties of kernel estimates of a regression function.* Journal of Statistical Planning and Inference, 4(1), 81-90.

[103] Grossman, S. and Stiglitz, J. (1980). *On the Impossibility of Informationally Efficient Markets.* American Economic Review, 70(3), 393-408.

[104] Guo, H., Wang, Z. and Yang, J. (2013). *Time-Varying Risk-Return Trade-off in the Stock Market.* Journal of Money, Credit and Banking, 45(4), 623-650.

[105] Guo, H. and Whitelaw, R. (2006). *Uncovering the Risk-Return Relation in the Stock Market.* Journal of Finance, 61(3), 1433-1463.

[106] Gut, A. (2005). *Probability: A Graduate Course.* New York: Springer.

[107] Han, H., Linton, O. B., Oka, T. and Whang, Y. J. (2014). *The Cross-Quantilogram: Measuring Quantile Dependence and Testing Directional Predictability between Time Series* (January 2014). Working paper available at SSRN: http://ssrn.com/abstract=2338468 or http://dx.doi.org/10.2139/ssrn.2338468.

[108] Hannan, E. J. (1970). *Multiple Time Series.* New York: John Wiley and Sons.

[109] Hannan, E. J. (1975). *Linear regression in continuous time.* Journal of the Australian Mathematical Society: Series A, 19(02), 146-159.

[110] Härdle, W. K. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

[111] Harvey, C. (2001). *The Specification of Conditional Expectations.* Journal of Empirical Finance, 8(5), 573-638.

[112] Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading.* New York: Oxford University Press.

[113] Hasbrouck, J. (2015). *High Frequency Quoting: Short-Term Volatility in Bids and Offers.* Working paper available at SSRN: http://ssrn.com/abstract=2237499. or http://dx.doi.org/10.2139/ssrn.2237499.

[114] Hayashi, T. and Yoshida, N. (2005). *On covariance estimation of non-synchronously observed diffusion processes.* Bernoulli, 11(2), 359-379.

[115] Hayashi, T. and Yoshida, N. (2008). *Asymptotic normality of a covariance estimator for nonsynchronously observed diffusion processes.* Annals of the Institute of Statistical Mathematics, 60(2), 367-406.

[116] Herrndorf, N. (1985). *A functional central limit theorem for strongly mixing sequences of random variables.* Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 69(4), 541-550.

[117] Hertzberg, D. (2010). *The Great Stock Myth: Why the market's rate of return—and your nest egg—may never recover.* The Atlantic. Working paper available at http://www.theatlantic.com/magazine/archive/2010/09/the-great-stock-myth/308178/.

[118] Hillman, R. and Salmon, M. (2007). *Intrinsic stationarity: investigating predictability in real-time forex transactions.* Journal of Financial Forecasting, 1(1), 3-43.

[119] Hirukawa, M. (2006). *A modified nonparametric prewhitened covariance estimator.* Journal of Time Series Analysis, 27(3), 441-476.

[120] Hong, S. Y., Lifshits, M. and Nazarov. A. (2016). *Small deviations in L_2-norm for Gaussian dependent sequences.* Electronic Communications in Probability, 21(41), 1-9.

[121] Hong, S. Y., Linton, O. B. (2017). *Nonparametric estimation of regression functions of infinite order with application to the risk-return tradeoff.* SSRN Working paper available at http://ssrn.com/abstract=2766822

[122] Hong, S. Y., Linton, O. B. and Zhang, H. Z. (2015). *An Investigation into Multivariate Variance Ratio Statistics and their Application to Stock Market Predictability.* Centre for Microdata Methods and Practice (CEMMAP) working paper CWP13/15 available at https://www.cemmap.ac.uk/uploads/cemmap/wps/cwp131515.pdf

[123] Hong, S. Y., Linton, O. B. and Zhang, H. Z. (2017). *An Investigation into Multivariate Variance Ratio Statistics and their Application to Stock Market Predictability.* Journal of Financial Econometrics, 15(2), 173-222.

[124] Hong, Y. (2000). *Generalized spectral tests for serial dependence.* Journal of the Royal Statistical Society: Series B, 62(3), 557-574.

[125] Hong, Y. and Lee, Y. J. (2005). *Generalized spectral tests for conditional mean models in time series with conditional heteroscedasticity of unknown form.* Review of Economic Studies, 72(2), 499-541.

[126] Hosking, J. R. M. (1981). *Equivalent forms of the multivariate portmanteau statistic.* Journal of the Royal Statistical Society: Series B, 43(2), 261-262.

[127] Ibragimov, I. A. (1962). *Some limit theorems for stationary processes.* Theory of Probability & Its Applications, 23, 291-300.

[128] Ibragimov, I. A. and Linnik, Y. B. (1971). *Independent and Stationary Sequences of Random variables.* Groningen: Wolters-Noordhoff.

[129] Jacod, J. (1994). *Limit of random measures associated with the increments of a Brownian semimartingale.* Unpublished paper: Laboratorie de Probabilities, Université de Paris VI.

[130] Jacod, J. (1997). *On continuous conditional Gaussian martingales and stable convergence in law.* Séminaire de probabilités de Strasbourg, 31, 232-246.

[131] Jacod, J., Li, Y., Mykland, P. A., Podolskij, M. and Vetter, M. (2009). *Microstructure noise in the continuous case: the pre-averaging approach*. Stochastic processes and their applications, 119(7), 2249-2276.

[132] Jacod, J. and Protter (1998). *Asymptotic error distributions for the Euler method for stochastic differential equations.* Annals of Probability, 26(1), 297-307.

[133] Jia, Q., Zhou, S. and Yin. F. (2003). *Kolmogorov entropy of global attractor for dissipative lattice dynamical systems.* Journal of Mathematical Physics, 44, 5804-5801.

[134] Jin, B., Wang, C., Bai, Z. D., Krishnan, N. K. and Harding, M. (2014) *Limiting spectral distribution of a symmetrized auto-cross covariance matrix.* Annals of Applied Probability, 24(3), 1199-1225.

[135] Johnstone, I.M. and A. Onatski (2015). *Testing in high-dimensional spiked models.* Available at https://arxiv.org/abs/1509.07269.

[136] Kahneman, D. and Tversky, A. (2000). *Choices, Values and Frames.* Cambridge: Cambridge University Press.

[137] Kalnina, I. and Linton, O. (2008). *Estimating quadratic variation consistently in the presence of endogenous and diurnal measurement error.* Journal of Econometrics, 147(1), 47-59.

[138] Kan, R. and Wang, X. (2010). *On the distribution of the sample autocorrelation coefficients.* Journal of Econometrics, 154(2), 101-121.

[139] Karamata, J. (1933). *Sur un mode de croissance regulire. Theoremes fondamentaux* (in French). Bulletin de la Societe Mathematique de France, 61, 55-62.

[140] Kara-Zaïtri, L., Laksaci, A., Rachdi, M. and Vieu, P. (2017). *Uniform in the smoothing parameter consistency results in functional regression.* Functional Statistics and Related Fields, 161-167, Springer.

[141] Karlin, S. and Taylor, H. (1975). *A first course in stochastic processes.* San Diego: Academic Press.

[142] Kim, M. J., Nelson, C. R. and Startz R. (1991). *Mean reversion in stock prices? A reappraisal of the empirical evidence.* Review of Economic Studies, 58(3), 515-528.

[143] King, M. (2016). *The End of Alchemy: Money, Banking and the Future of the Global Economy.* New York: Norton & Company.

[144] Kirby, C. (1998). *The restrictions on predictability implied by rational asset pricing models.* Review of Financial Studies, 11(2), 343-382.

[145] Kudraszow, N. L. and Vieu, P. (2013). *Uniform consistency of kNN regressors for functional variables.* Statistics and Probability Letters, 83(8), 1863-1870.

[146] Lettau, M. and Ludvigson, S. (2010). *Measuring and modeling variation in the risk-return tradeoff.* In: Ait-Sahalia, Y., Hansen, L. (Eds.), Handbook of Financial Econometrics. Amsterdam: North-Holland.

[147] Li, W. V. (2012). *Small value probabilities in analysis and mathematical physics.* Presented at the Arizona School of Analysis and Mathematical Physics, Tucson, Arizona, United States in March 15, 2012. Retrieved from the link `http://math.arizona.edu/~mathphys/school_2012/WenboLi.pdf`.

[148] Li, W. V. and Shao, Q. M. (2001). *Gaussian processes: inequalities, small ball probabilities and applications.* Handbook of Statistics, 19, 533-598.

[149] Linton, O. and Perron, B. (2003). *The Shape of the Risk Premium: Evidence From a Semiparametric Generalized Autoregressive Conditional Heteroscedasticity Model.* Journal of Business and Economic Statistics, 21(3). 354-367.

[150] Linton, O. B. and Sancetta, A. (2009). *Consistent estimation of a general nonparametric regression function in time series.* Journal of Econometrics, 152(1), 70-78.

[151] Liu, W. and Wu, W. B. (2010). *Asymptotics of spectral density estimates.* Econometric Theory, 26(4), 1218-1245.

[152] Lo, A. W. (2004). *The adaptive markets hypothesis.* Journal of Portfolio Management, 30(5), 15-29.

[153] Lo, A. W. (2005). *Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis.* Journal of Investment Consulting, 7(2), 21-44.

[154] Lo, A. W. and MacKinlay, A. C. (1999). *A Non-Random Walk Down Wall Street.* New Jersey: Princeton University Press.

[155] Lo, A. W. and MacKinlay, A. C. (1988). *Stock market prices do not follow random walks: Evidence from a simple specification test.* Review of Financial Studies, 1(1), 41-66.

[156] Lo, A. W. and MacKinlay, A. C. (1989). *The size and power of the variance ratio test in finite samples: A Monte Carlo investigation.* Journal of Econometrics, 40, 203-238.

[157] Lo, A. W. and MacKinlay, A. C. (1990). *When are contrarian profits due to stock market overreaction?*. Review of Financial studies, 3(2), 175-205.

[158] Lobato, I. N. (2001). *Testing that a dependent process is uncorrelated. Journal of the American Statistical Association*, 96(455), 1066-1076.

[159] Lu, Z. (2001). *Asymptotic normality of kernel density estimators under dependence.* Annals of the Institute of Statistical Mathematics, 53(3), 447-468.

[160] Ludvigson, S. and Ng, S. (2007) *The Empirical Risk-Return Relation: A Factor Analysis Approach*. Journal of Financial Economics, 83(1), 171-222.

[161] Lundblad, C. (2005). *The Risk-Return Tradeoff in the Long Run: 1836-2003*. Journal of Financial Economics, 85(1), 123-150.

[162] Magdalinos (2014). Personal communication.

[163] Magnus, J. R. and Neudecker, H. (1980). *The elimination matrix: some lemmas and applications*. SIAM Journal on Algebraic Discrete Methods, 1(4), 422-449.

[164] Malkiel, B. G. (2015). *A random walk down wall street*. New York: Norton & Company.

[165] Malliavin, P. and Mancino, M.E. (2002). *Fourier series method for measurement of multivariate volatilities*. Finance and Stochastics, 6(1), 49-61.

[166] Malliavin, P. and Mancino, M. E. (2009). *A Fourier transform method for nonparametric estimation of multivariate volatility*. Annals of Statistics, 37(4), 1983-2010.

[167] Mammen E. (1991) *Nonparametric Curve Estimation and Simple Curve Characteristics*. In: Roussas G. (eds) Nonparametric Functional Estimation and Related Topics. NATO ASI Series (Series C: Mathematical and Physical Sciences), vol 335. Springer, Dordrecht.

[168] Mancino, M. E. and Sanfelici, S. (2008). *Robustness of Fourier estimator of integrated volatility in the presence of microstructure noise*. Computational Statistics and Data Analysis, 52(6), 2966-2989.

[169] Mas, A. (2012). *Lower bound in regression for functional data by small ball probability representation in Hilbert space*, Electronic Journal of Statistics, 6, 1745-1778.

[170] Masry, E. (2005). *Nonparametric regression estimation for dependent functional data: asymptotic normality*. Stochastic Processes and their Applications, 115(1), 155-177.

[171] Masry, E. and Fan, J. (1997). *Local polynomial estimation of regression function for mixing processes*. Scandinavian Journal of Statistics, 24(2), 1965-1979.

[172] Mehra, R. (2012). *Consumption-Based Asset PRicing Models'*. Annual Review of Financial Economics, 4(1), 385-409.

[173] Merlevède, F., Peligrad, M., Rio, E. (2011). *A Bernstein type inequality and moderate deviations for weakly dependent sequences*. Probability Theory and Related Fields, 151(3-4), 435-474.

[174] Merton, R. C. (1973). *An Intertemporal Capital Asset Pricing Model*. Econometrica, 41(5), 867-887.

[175] Mikosch, T. and Stărică, C. (2000). *Limit theory for the sample autocorrelations and extremes of a GARCH (1, 1) process*. Annals of Statistics, 28(5), 1427-1451.

[176] Muth, J. F. (1960). *Optimal properties of exponentially weighted forecasts*. Journal of the American Statistical Association, 55(290), 299-306.

[177] Mykland, P. A. and Zhang, L. (2006). *ANOVA for diffusions and Itô processes*. Annals of Statistics, 34(4), 1931-1963.

[178] Nadaraya, E. A. (1964). *On estimating regression*. Theory of Probability & Its Applications, 9(1), 141-142.

[179] Nadaraya, E. A. (1970). *Remarks on non-parametric estimates for density functions and regression curves*. Theory of Probability & Its Applications, 15(1), 134-137.

[180] Nelson, D. B. (1991). *Conditional Heteroscedasticity in Asset Returns: a New Approach*. Econometrica, 59(2), 347-370.

[181] Newey, W. K. and West, K. D. (1987). *Hypothesis testing with efficient method of moments estimation*. International Economic Review, 28(3), 777-787.

[182] O'Hara, M. and Ye, M. (2009). *Is fragmentation harming market quality?*. Journal of Financial Economics, 100(3), 459-474.

[183] Olver, F. W. J., Lozier, D. W., Boisvert, R. F. and Clark, C. W. (2010). *NIST Handbook of Mathematical Functions*. New York: Cambridge University Press.

[184] Pagan, A. R. and Hong, Y. S. (1990). *Nonparametric Estimation and the Risk Premium*. In W. Barnett, J. Powell, and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics*. Cambridge University Press.

[185] Pagan, A. R. and Ullah, A. (1988). *The econometric analysis of models with risk terms*. Journal of Applied Econometrics, 3(2), 87-105.

[186] Pagan, A. R. and Ullah, A. (1999). *Nonparametric econometrics.* Cambridge: Cambridge University Press.

[187] Park, S. (2011). *Consistent Estimator of Ex-post Covariation of Discretely Observed Diffusion Processes and Its Application to High Frequency Financial Time Series.* PhD thesis, The London School of Economics and Political Science (LSE).

[188] Park, S., Hong, S. Y. and Linton, O. B. (2016). *Estimating the quadratic covariation matrix for asynchronously observed high frequency stock returns corrupted by additive measurement error.* Journal of Econometrics, 191(2), 325-347.

[189] Park, S. and Linton, O. (2012). *Realized Volatility: Theory and Applications.* In: Handbook of Volatility Models and Their Applications, Wiley, eds. Bauwens, L., Hafner, C. and Sebastien, L. London, 319-345.

[190] Parzen, E. (1957). *On consistent estimates of the spectrum of a stationary time series.* Annals of Mathematical Statistics, 28(2), 329-348.

[191] Parzen, E. (1962). *On estimation of a probability density function and mode.* Annals of Mathematical Statistics, 33(3), 1065-1076.

[192] Parzen, E. (1967). *The role of spectral analysis in time series analysis.* Review of the International Statistical Institute, 35(2), 125-141.

[193] Pástor, L., Sinha, M. and Swaminathan, B. (2008). *Estimating the Intertemporal Risk-Return Tradeoff Using the Implied Cost of Capital.* Journal of Finance, 63(6), 2859-2897.

[194] de la Peña, V. H., Lai, T. L. and Shao, Q. M. (2009). *Self-normalized processes: Limit theory and Statistical Applications.* New York: Springer.

[195] Pesaran, M. H. and Timmermann, A. (2007). *Selection of estimation window in the presence of breaks.* Journal of Econometrics, 137(1), 134-161.

[196] Phillips, P. C. B. (2015). *Halbert White Jr. Memorial JFEC Lecture: Pitfalls and Possibilities in Predictive Regression.* Journal of Financial Econometrics, 13(3), 521-555.

[197] Phillips, P. C. B. and Jin, S. (2014). *Testing the Martingale Hypothesis.* Journal of Business & Economic Statistics, 32(4), 537-554.

[198] Phillips, P. C. B. and Magdalinos, T. (2007). *Limit theory for moderate deviations from a unit root.* Journal of Econometrics, 136(1), 115-130.

[199] Phillips, P. C. B. and Park, J. Y. (1998). *Nonstationary density estimation and kernel autoregression*. Cowles Foundation discussion paper.

[200] Phillips, P. C. B., Shi, S. P. and Yu, J. (2012). *Testing for multiple bubbles*. Cowles Foundation Discussion Paper no. 1843.

[201] Phillips, P. C. B. and Solo, V. (1992). *Asymptotics for Linear Processes*. Annals of Statistics, 20(2), 971-1001.

[202] Phillips, P. C. B. and Yu, J. (2011). *Dating the timeline of financial bubbles during the subprime crisis*. Quantitative Economics, 2(3), 455-291.

[203] Pierce, D. A. and Haugh, L. D. (1977). *Causality in temporal systems: Characterization and a survey*. Journal of Econometrics, 5(3), 265-293.

[204] Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer-Verlag.

[205] Poterba, J. M. and Summers, L. H. (1988). *Mean reversion in stock prices: Evidence and Implications*. Journal of Financial Economics, 22(1), 27-59.

[206] Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. New York: Springer.

[207] Subba Rao, T. and Gabr, M. M. (1984). *An introduction to bispectral analysis and bilinear time series models*. In: Lecture Notes in Statistics. Springer-Verlag, New York.

[208] Richardson, M. and Stock, J. H. (1989). *Drawing inferences from statistics based on multiyear asset returns*. Journal of Financial Economics, 25(2), 323-348.

[209] Rio, E. (2000). *Theorie asymptotique des processus alé atoires faiblement dependants* (in French). Berlin: Springer Verlag.

[210] Robinson, P. M. (1975). *Continuous time regressions with discrete data*. Annals of Statistics, 3(3), 688-697.

[211] Robinson, P. M. (1983). *Nonparametric estimators for time series*. Journal of Time Series Analysis, 4(3), 185-207.

[212] Rosenblatt, M. (1956). *A central limit theorem and a strong mixing condition*. Proceedings of the National Academy of Sciences, 42(1), 43-47.

[213] Rosenblatt, M. (1985). *Stationary Sequences and Random Fields*, Boston: Birkhäuser.

[214] Ross, S. A. (2002). *A Neoclassical Look at Behavioral Finance; Closed End Funds.* (The Princeton Lectures in Finance III).

[215] Roussas, G. G. (1989). *Consistent regression estimation with fixed design points under dependence conditions.* Statistics & Probability Letters, 8(1), 41-50.

[216] Roussas, G. G. (1990). *Nonparametric regression estimation under mixing conditions.* Stochastic Processes and Their Applications, 36(1), 107-116.

[217] Shao, X. and Wu, W. B. (2007). *Asymptotic spectral theory for nonlinear time series.* Annals of Statistics, 35(4), 1773-1801.

[218] Schuster, E. F. (1972). *Joint asymptotic distribution of the estimated regression function at a finite number of distinct points.* Annals of Mathematical Statistics, 43(1), 84-88.

[219] Scruggs, J. (1998). *Resolving the Puzzling Intertemporal Relation Between the Market Risk Premium and the Conditional Market variance: A Two-Factor Approach.* Journal of Finance, 53(2), 575-603.

[220] Scruggs, J. and Glabadanidis, P. (2003). *Risk Premia and the Dynamic covariance Between Stock and Bond Returns.* Journal of Financial and Quantitative Analysis, 38(2), 295-316.

[221] Sheppard, K. (2013). *Measuring Market Speed* (March 2013). Working paper available at http://www.kevinsheppard.com/images/f/fa/MeasuringMarketSpeed.pdf.

[222] Shiller, R. A. (2013). *Speculative Asset prices.* Nobel prize lecture.

[223] Smith, D. R. and Whitelaw, R. (2009). *Time-varying risk aversion and the risk-return relation.* SSRN Working paper available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1663542

[224] Solnik, B. (1991). *International investments.* Reading, MA: Addison Wesley.

[225] Stone, C. (1980). *Optimal rates of convergence for nonparametric estimators.* Annals of Statistics, 8(6), 1348-1360.

[226] Stone, C. (1982). *Optimal global rates of convergence for nonparametric regression.* Annals of Statistics, 10(4), 1040-1053.

[227] Sytaya, G. N. (1974). *On certain asymptotic representations for a Gaussian measure in Hilbert space* (in Russian). Theory of Random Process, 2, 93-104.

[228] Szroeter, J. (1978). *Generalized variance-ratio tests for serial correlation in multivariate regression models.* Journal of Econometrics, 8(1), 47-59.

[229] Taylor, S. J. (2005). *Asset Price Dynamics, Volatility, and Prediction.* New Jersey: Princeton University Press.

[230] Timmerman, A. (2008). *Elusive Return Predictability.* International Journal of Forecasting, 24(1), 1-18.

[231] Veronesi, P. (2000): "How Does Information Quality Affect Stock Returns?," The Journal of Finance, 55(2), 807–837.

[232] Watson, G. S. (1964). *Smooth regression analysis.* Sankhya Series A, 26(4), 359-372.

[233] Whang, Y. J. and Kim, J. (2003). *A multiple variance ratio test using subsampling.* Economics Letters, 79(2), 225-230.

[234] White, H. (1984). *Asymptotic theory for econometricians.* New York: Academic Press.

[235] White, H., (2000), *A Reality Check for Data Snooping*, Econometrica 68, 1097-1126

[236] White, H., and I Domowitz (1984). *Nonlinear Regression with Dependent Observations.* Econometrica 52, 143-162.

[237] Whitelaw, R. (1994). *Time variations and covariations in the Expectation and Volatility of Stock Market Returns*, Journal of Finance, 49(2), 515-541.

[238] Whitelaw, R. F. (2000). *tock Market Risk and Return: An Equilibrium Approach*, Review of Financial Studies, 13(3), 521-547.

[239] Wright, J. H. (2000). *Alternative variance-ratio tests using ranks and signs.* Journal of Business & Economic Statistics, 18(1), 1-9.

[240] Wu, W. B. (2011). *Asymptotic theory for stationary processes.* Statistics and Its Interface, 4, 207-226.

[241] Xiao, Z. and Linton, O. (2002). *A nonparametric prewhitened covariance estimator.* Journal of Time Series Analysis, 23(2), 215-250.

[242] Yaracos, Y. G. (1985). *Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy.* Annals of Statistics, 13(2), 768-774.

[243] Zhang, L. (2006). *Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach.* Bernoulli, 12(6), 1019-1043.

[244] Zhang, L. (2011). *Estimating covariation: Epps effect, microstructure noise.* Journal of Econometrics, 160(1), 33-47.

[245] Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005). *A tale of two time scales.* Journal of the American Statistical Association, 100(472), 1394-1411.

[246] Zhang, X., King, M. L. and Hyndman, R. J. (2006). *A Bayesian approach to bandwidth selection for multivariate kernel density estimation.* Computational Statistics and Data Analysis, 50(11), 3009-3031.

[247] Zivot, E. (2005). *Analysis of High Frequency Financial Data: Models, Methods and Software. Part II: Modeling and Forecasting Realized Variance Measures.* Unpublished Manuscript available at https://faculty.washington.edu/ezivot/realizedvariance.pdf

[248] Zolotarev, V. M. (1986). *Asymptotic behavior of the Gaussian measure in ℓ_2.* Journal of Soviet Mathematics, 35(2), 2330-2334.